Pier Luigi Conti
Daniela Marella

Campionamento da popolazioni finite

Il disegno campionario

UNITEXT

Springer

EXTRA MATERIALS
extras.springer.com

Alla memoria del nostro caro Maestro, Gianni Tranquilli

Pier Luigi Conti • Daniela Marella

Campionamento da popolazioni finite

Il disegno campionario

Springer

Pier Luigi Conti
Dipartimento di Scienze Statistiche
Sapienza Università di Roma

Daniela Marella
Dipartimento di Scienze dell'Educazione
Università Roma Tre

Contenuti integrativi sono consultabili su http://extras.springer.com
Password: 978-88-470-2576-9

UNITEXT – Collana di Statistica e Probabilità Applicata
ISSN 2038-5714 e-ISSN 2038-5765

Additional material to this book can be downloaded from http://extras.springer.com
ISBN 978-88-470-2576-9 ISBN 978-88-470-2577-6 (eBook)
DOI 10.1007/978-88-470-2577-6

Springer Milan Heidelberg New York Dordrecht London

9 8 7 6 5 4 3 2 1

Layout copertina: Beatrice ꝺ, Milano

Impaginazione: PTP-Berlin, Protago TEX-Production GmbH, Germany (www.ptp-berlin.eu)
Stampa: Grafiche Porpora, Segrate (MI)

Springer-Verlag Italia S.r.l., Via Decembrio 28, I-20137 Milano
Springer-Verlag fa parte di Springer Science+Business Media (www.springer.com)

Prefazione

Viviamo nella società dell'informazione. Non passa giorno senza che questa frase ci sia ricordata da giornali, televisioni, Internet. Non passa giorno senza che veniamo bombardati di cifre riguardanti gli elementi più disparati del nostro sistema sociale: andamento dei prezzi, livello di disoccupazione, gradimento nei confronti di questo o quel partito politico. Non passa giorno senza che ci siano forniti i risultati di "sondaggi" riguardanti i più vari aspetti del nostro vivere sociale.

Questa formidabile mole di numeri che ci vengono forniti in dosi sempre più massicce, però, genera spesso scetticismo. Come stabilire quando l'informazione fornita può definirsi "corretta"? Come evitare di fornire numeri molto lontani dalla realtà che si pretende di rappresentare, e su cui si pretende di informare i cittadini?

Questo libro si occupa della corretta acquisizione e dell'uso efficiente di un tipo molto importante di informazione: l'*informazione statistica*. Con una certa semplificazione, e con qualche imprecisione, l'acquisizione dell'informazione statistica ha luogo tramite l'*osservazione* di una o più caratteristiche di interesse (*status* occupazionale, partito politico per cui si intende votare, etc.) sulle unità di una *popolazione* di riferimento.

Un elemento critico del processo di acquisizione dei dati è costituito dal fatto che molto spesso le popolazioni di riferimento sono composte da un numero molto elevato di unità, la cui osservazione completa richiederebbe costi e tempi proibitivi. Per tale ragione si ricorre a processi di acquisizione dei dati basati sull'osservazione di una parte delle unità della popolazione, la quale costituisce un *campione*.

L'idea-guida del presente libro è sostanzialmente una: sia il processo di selezione del campione che l'uso dei dati corrispondenti devono essere volti ad ottenere la massima *efficienza*. È questo obiettivo a spingere la trattazione verso regole di selezione del campione di tipo "scientifico", fondate sul calcolo delle probabilità. Solo in questo caso, infatti, è possibile studiare in modo corretto cosa si intenda per "uso efficiente dei dati statistici".

Questo volume è dedicato in larga misura agli aspetti di base del campionamento da popolazioni finite. Vengono stabilite le basi logiche del campionamento e ne vengono studiati i principali sviluppi elementari. L'unica fonte di errore statistico è dovuta al fatto che non si osservano tutte le unità della popolazione, ma solo una parte. Ad esso seguirà un successivo volume dedicato a tematiche più avanzate, generate (prevalentemente ma non esclusivamente) sia dal fatto che a volte può essere impossibile osservare alcune delle unità del campione e/o è possibile effettuare solo osservazioni affette da errori, sia dal fatto che spesso l'interesse verte sulla costruzione di *modelli statistici* per i dati concretamente osservati.

L'approccio espositivo seguito consiste nel procedere dal particolare al generale, partendo da aspetti il più possibile elementari, che vengono poi "complicati" per renderli via via più aderenti alle concrete situazioni applicative. Molti ragionamenti vengono prima basati su esempi concreti e poi su aspetti teorici che, se forniti in prima battuta, potrebbero rendere un po' ostica la comprensione di aspetti chiave. A questo proposito si sottolinea l'importanza degli esempi basati su dati (tutti disponibili sulla pagina web http://extras.springer.com), i quali non sono solo un utile complemento alla teoria, ma una vera e propria chiave di accesso al "ragionamento statistico". Lo stesso ruolo è anche svolto da molti degli esercizi proposti.

Il libro si rivolge sia a studenti universitari di corsi di laurea con una robusta componente quantitativa (Scienze Statistiche, Economia, etc.), sia a ricercatori in campo economico o sociale che utilizzano il metodo statistico. Il livello complessivo della trattazione è, per quanto possibile, elementare. L'unico pre-requisito richiesto per accedere a gran parte degli argomenti trattati è un corso elementare di statistica con elementi introduttivi di calcolo delle probabilità e inferenza statistica, come quelli impartiti nelle lauree triennali di Economia e Scienze Politiche. Non strettamente necessario, anche se utile, è un corso elementare di matematica, modellato sui corsi di matematica generale dei corsi di laurea in Economia.

Questo volume è prevalentemente destinato alle lauree triennali. Nella nostra esperienza, i Capitoli 1–7, la Sezione 8.6, i Capitoli 9–11 costituiscono il materiale per un corso di 32–36 ore in una laurea triennale di taglio statistico. L'aggiunta del Capitolo 12, di ampie parti del Capitolo 14 (lo stimatore di Horvitz-Thompson, in sostanza) e di alcune sezioni del Capitolo 15 (disegni di Poisson, di Poisson condizionato e di Pareto) copre un corso *standard* di campionamento di 48 ore, sempre per una laurea triennale di taglio statistico. Lo stesso materiale è stato anche usato, a vari livelli, in corsi di laurea e di Master in Facoltà di Economia. Alcune parti dei Capitoli 14 (stimatori di tipo calibrazione), 15 (disegno bilanciato) e gran parte del materiale sugli errori non campionari e sui modelli statistici per dati campionari (oggetto di trattazione in un successivo volume) sono invece destinati alle lauree specialistiche.

Le parti (sezioni o interi capitoli) con asterisco sono più avanzate, e non vengono utilizzate, in generale, nell'ambito delle lauree triennali. Alcune di

esse sono usate in lauree specialistiche, mentre altre sono state incluse in quanto di diretto interesse applicativo.

Molti degli esempi proposti nel volume si basano su *dataset* disponibili presso la pagina web http://extras.springer.com.

Nello scrivere il presente libro abbiamo contratto parecchi debiti di gratitudine. Il Dott. Mauro Scanu ha letto l'intero volume, indicando parecchi errori e fornendo utili suggerimenti. I Proff. Ludovico Piccinato, Marco Riani, Paola Vicard hanno letto ampie parti del volume, fornendo spunti e considerazioni di grande interesse. Utili suggerimenti sono stati anche forniti dal Prof. Francesco Battaglia. Naturalmente, di errori e omissioni residui sono responsabili i soli autori.

Roma, febbraio 2012

Pier Luigi Conti
Daniela Marella

Indice

1

Aspetti generali sul campionamento da popolazioni finite

1.1 Rilevazioni censuarie e rilevazioni campionarie

La necessità di informazioni statistiche sempre più accurate e disponibili in tempi rapidi costituisce indubbiamente uno degli aspetti salienti delle società moderne. Tali informazioni sono molto spesso acquisibili solo mediante *rilevazioni statistiche*, che consistono (almeno in prima approssimazione) nelle attività di raccolta ed elaborazione di dati statistici riguardanti specifici insiemi di elementi, detti "popolazioni finite". Gli esempi in proposito sono numerosissimi. Ad es., il nostro obiettivo potrebbe essere quello di ottenere informazioni sulle aziende operanti in uno o più settori di attività economica e sul loro fatturato, sugli individui di una data comunità (nazionale, regionale, etc.) e sul loro *status* lavorativo (occupati, disoccupati, etc.), sulle famiglie residenti in una data area geografica e su aspetti legati alle loro abitudini di consumo e ai loro redditi, sui cittadini aventi diritti politici e sulla loro disponibilità a votare un certo partito politico. Solo se resi disponibili con tempestività, tali dati possono soddisfare le specifiche esigenze, sia conoscitive che decisionali, di istituzioni pubbliche e organismi privati.

La risposta a tali esigenze molto di rado può venire da rilevazioni censuarie, in cui si osservano tutti gli elementi di una data popolazione. In genere le popolazioni di interesse sono costituite da numerosi elementi, per cui la sola raccolta dei dati (senza contare le successive fasi di elaborazione) richiederebbe un gran dispiego di mezzi, disponibilità finanziarie assai ingenti, e tempi di esecuzione inevitabilmente lunghi.

Nei casi di censimenti demografici e socio-economici riguardanti un'intera nazione, le rilevazioni censuarie non possono che essere intraprese pubblicamente dal rispettivo stato, e condotte istituzionalmente dal relativo ufficio statistico nazionale. Tali rilevazioni totali sono svolte, in ogni paese, con cadenza temporale regolare (in genere decennale), e tendono a rimanere circoscritte alle sole informazioni socio-demografiche ed economiche di interesse generale.

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_1,

Altre informazioni statistiche totali di carattere pubblico sono regolarmente raccolte come indiretto sottoprodotto di attività di registrazione e di controllo della Pubblica Amministrazione (statistiche amministrative). I dati statistici che ne derivano sono spesso del tutto insoddisfacenti sia sotto il profilo della tempestività, in quanto in genere si rendono disponibili quando non sono più utili, che sotto quello della qualità del (sotto-)prodotto offerto. Se soddisfano condizioni di attendibilità, tali statistiche amministrative possono essere utilizzate per analisi storiche retrospettive e per elaborare modelli previsionali da impiegare per prevedere il presente sulla base del passato, anche se sul presente sarebbe sempre meglio indagare direttamente tramite rilevazioni statistiche.

Quanto sopra osservato motiva il ricorso a *rilevazioni campionarie* (o *parziali*), in cui si osserva solo una parte (in genere piccola) della popolazione oggetto di interesse. La parte della popolazione osservata è denominata *campione*, che è inteso come un *rappresentante* della popolazione complessiva.

Rispetto alle rilevazioni totali, quelle parziali presentano alcuni fondamentali vantaggi.

- Hanno tempi di esecuzione assai più rapidi, e quindi permettono di disporre con relativa tempestività dei dati statistici e delle relative elaborazioni.
- Hanno costi assai più contenuti. Le rilevazioni censuarie su larga scala impegnano in genere un gran numero di rilevatori, e quindi hanno costi sostenibili quasi soltanto da enti pubblici ad esse preposti. Per questa ragione molte rilevazioni censuarie sono state dismesse e sostituite da rilevazioni parziali. Ad es., il censimento dell'industria e servizi 2010 è consistito in realtà in una rilevazione parziale. Nessun paese dell'Unione Europea conduce più "vere" rilevazioni censuarie di fenomeni di natura economica.
- Forniscono dati in genere *più accurati*. Le rilevazioni campionarie, osservando soltanto una frazione ridotta della popolazione, possono utilizzare pochi rilevatori molto ben addestrati, e forniscono risultati affidabili. Per contro, le indagini censuarie, avendo bisogno di molti rilevatori, tenderanno spesso ad usare anche personale poco addestrato e non qualificato, che probabilmente rileverà i dati in maniera poco accurata. Inoltre, le indagini censuarie, trattando grosse moli di dati, sono più esposte ad errori di trascrizione, codifica, etc., che abbassano ulteriormente la qualità dei dati prodotti. Per questa ragione i pochi censimenti effettuati prevedono spesso una rilevazione parziale *a posteriori* per il controllo della qualità dei dati.

In alcuni casi è la natura stessa dell'indagine che determina il tipo di rilevazione da utilizzare. Ad esempio, nel controllo statistico della qualità è la natura del processo di misurazione che, comportando la distruzione dell'unità che si osserva, obbliga il ricorso ad una rilevazione campionaria (ad es. durata di accensione di una lampadina).

Le rilevazioni campionarie, d'altra parte, hanno un inconveniente di rilievo: il campione (parte di popolazione osservata) potrebbe essere un pessimo rappresentante della popolazione totale, e quindi fornire risultati di scarsa

utilità, o addirittura controproducenti. 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. In ogni caso, a fronte dello svantaggio dianzi menzionato, le rilevazioni campionarie presentano tanti e tali vantaggi rispetto a quelle censuarie che la pratica del campionamento è ormai diffusissima nei più svariati settori, e si può affermare che la gran parte delle rilevazioni statistiche comunemente effettuate sono di tipo campionario.

1.2 Linee metodologiche di una rilevazione statistica

Qui di seguito sono brevemente delineate le linee metodologiche essenziali di una rilevazione statistica (censuaria o campionaria).

- Una rilevazione riguarda un insieme finito di *elementi individuali* (detti anche *unità di osservazione*, o *unità elementari*), i quali costituiscono una *popolazione finita*. Gli elementi che compongono una popolazione sono ben definite entità, fisicamente esistenti.
- Sulle unità elementari che costituiscono una popolazione sono definiti uno o più *caratteri* oggetto di interesse. In corrispondenza di ogni unità elementare ciascun carattere si manifesta con una determinata *modalità*. Una rilevazione statistica ha in genere l'obiettivo di studiare il "modo di manifestarsi" del(i) carattere(i) di interesse sulla popolazione di riferimento. Molto spesso le modalità di un carattere sono riassunte mediante opportune sintesi descrittive (come media, varianza, o altro), le quali enucleano aspetti particolarmente significativi del modo in cui uno o più caratteri si manifestano in una popolazione. Tali sintesi descrittive costituiscono dei *parametri statistici di interesse*. Un obiettivo importantissimo di una rilevazione statistica è quello di ottenere informazioni su tali parametri.
- L'accesso agli elementi (unità elementari) che costituiscono una popolazione è realizzato tramite una *lista*, che può essere vista come un "meccanismo" che ad ogni unità elementare della popolazione associa una *unità di campionamento* della lista. Non necessariamente le unità elementari coincidono con le unità di campionamento.
- Dalla lista di unità viene selezionato un suo sottoinsieme, denominato *campione*. Un ingrediente di fondamentale importanza, ovviamente, è rappresentato dalla *regola* con cui il campione è selezionato dalla popolazione. In particolare, nel seguito si concentrerà l'attenzione sulle regole di selezione di tipo *probabilistico*, le quali si basano sull'idea di selezionare il campione in accordo con una legge di probabilità determinata in fase di progettazione della rilevazione. La regola di selezione del campione è in genere denominata *disegno* (o *piano*) *campionario*.
- Alle unità di campionamento che formano il campione effettivamente selezionato corrispondono alcune delle unità di osservazione (elementi) della popolazione. Queste ultime costituiscono gli elementi della popolazione effettivamente campionati. Come già rimarcato, il ruolo primario delle unità

di campionamento è proprio quello di permettere l'accesso alle unità di osservazione della popolazione. Per gli elementi (unità di osservazione) selezionati nel campione, si *osservano* le modalità dei caratteri oggetto di interesse. L'osservazione di tali modalità è realizzata tramite un qualche *processo di misurazione*. Nei casi più comuni lo *strumento di misura* usato in questo processo è il *questionario*, ossia un elenco di domande alle quali gli individui selezionati devono fornire risposta. Un aspetto assai delicato ed importante è la messa a punto del questionario, in genere realizzata tramite un'indagine pilota. Le modalità osservate degli individui del campione vengono registrate, e costituiscono i *dati campionari*.

– I dati campionari vengono usati per calcolare delle *stime* dei parametri di interesse (medie, varianze, coefficienti di correlazione e di regressione, e altro ancora).

Una rilevazione censuaria può essere vista come caso particolare di rilevazione campionaria, in cui si selezionano (e osservano) tutte le unità elementari della popolazione.

Per illustrare meglio quanto sopra esposto, è opportuno ricorrere ad un esempio: la rilevazione delle forze di lavoro effettuata dall'Istituto Nazionale di Statistica (ISTAT). La rilevazione è di tipo campionario, ed ha cadenza trimestrale. La *popolazione* di riferimento è costituita da tutti gli individui residenti in Italia, inclusi quelli temporaneamente residenti all'estero, ed esclusi quelli che vivono abitualmente all'estero e i membri permanenti delle convivenze (ospizi, caserme, istituti religiosi, brefotrofi, etc.). Tali individui sono quindi le *unità elementari* (di osservazione) che formano la popolazione.

Degli individui componenti la popolazione interessano principalmente *caratteristiche* quali l'appartenenza alle forze di lavoro, lo *status* occupazionale, e (molto) altro ancora. Gli individui vengono distinti in due categorie: gli appartenenti alle forze di lavoro, distinti a loro volta in (a) occupati e, (b) disoccupati, e (c) "non appartenenti alle forze di lavoro". Gli *occupati* sono tutti coloro che, oltre ad avere un'età di almeno 15 anni, possiedono una delle seguenti caratteristiche:

a1. hanno effettuato una o più ore lavorative retribuite (o non retribuite se prestate in un'impresa familiare) nella settimana di riferimento della rilevazione;
a2. hanno un'attività lavorativa, anche se durante la settimana di riferimento della rilevazione non hanno lavorato.

Vengono invece classificati come disoccupati tutti gli individui (di almeno 15 anni di età) non occupati che sono in cerca di occupazione, in quanto in possesso di una delle seguenti caratteristiche:

b1. hanno effettuato almeno un'azione di ricerca "attiva" di lavoro nelle quattro settimane che precedono la rilevazione e sono immediatamente disponibili (entro due settimane) ad accettare un lavoro, qualora venga loro offerto;

b2. hanno già trovato un lavoro che inizierà nelle settimane successive alla rilevazione (non oltre tre mesi).

Gli individui appartenenti alle categorie (a1), (a2), (b1), (b2) costituiscono le *forze di lavoro*, mentre i rimanenti formano le *non forze di lavoro*.

I *parametri di interesse* sono in questo caso grandezze quali l'ammontare (numero di individui) delle forze di lavoro, il tasso di disoccupazione (uguale al rapporto tra disoccupati e totale delle forze di lavoro), e (molto) altro ancora.

L'*accesso* alla popolazione degli individui avviene tramite la *lista delle famiglie*, che costituiscono quindi le *unità di campionamento*. Si noti come le unità di campionamento, in questo caso, siano diverse da quelle elementari di osservazione. In effetti, ogni famiglia è un aggregato (un *grappolo* in termini tecnici) di unità elementari.

Il *disegno campionario* (regola di selezione delle famiglie) è di tipo probabilistico. L'idea di base è piuttosto semplice, e consiste in una procedura a *due stadi*. Si considerano in primo luogo le province italiane. Per ogni provincia si determina, con procedura *ad hoc*, una soglia dimensionale demografica. I comuni con una popolazione residente al di sopra della soglia vengono denominati "auto-rappresentativi", mentre quelli con una popolazione al di sotto della soglia sono "non auto-rappresentativi". I comuni di ogni provincia che non sono auto-rappresentativi, inoltre, vengono suddivisi in gruppi (*strati*) omogenei rispetto al peso demografico. La procedura di selezione delle famiglie è a due stadi, e può essere sintetizzata come segue:

- al primo stadio, per ogni provincia si selezionano: (*i*) tutti i comuni auto-rappresentativi; (*ii*) da ogni strato (di comuni non auto-rappresentativi) in cui la provincia è suddivisa, due comuni non-autorappresentativi;
- dalle liste anagrafiche dei comuni scelti al primo stadio, si seleziona un campione di famiglie.

Per gli individui che formano le famiglie selezionate nel campione si *osservano* le *modalità dei caratteri di interesse*. L'osservazione avviene mediante la somministrazione di un *questionario*, il quale costituisce un vero e proprio strumento di misura per il fenomeno oggetto di indagine. Per ogni individuo a cui è sottoposto il questionario si registrano le caratteristiche di appartenenza o meno alle forze di lavoro, status occupazionale, e (molto) altro ancora. In questo modo, vengono costruiti i *dati campionari* prodotti dalla rilevazione.

I dati campionari vengono usati per produrre *stime*, a vari livelli territoriali (provinciale, regionale, nazionale) di *parametri* quali la consistenza delle forze di lavoro, il tasso di disoccupazione, etc.

La rilevazione delle forze di lavoro è ripetuta con cadenza trimestrale. Tuttavia, in ogni trimestre *non* viene ripetuto l'intero processo di selezione di comuni e famiglie. Infatti ogni famiglia selezionata viene osservata per due trimestri consecutivi, esce temporaneamente dal campione per altri due trimestri, e viene di nuovo osservata nei due trimestri successivi. I comuni auto-rappresentativi sono sempre presenti nel campione, mentre quelli non auto-

rappresentativi vengono sostituiti quando non sono più in grado di fornire nuove famiglie al campione.

1.3 Popolazioni, etichette, modalità etichettate

Consideriamo una popolazione finita $\mathcal{U} = \{u_1, \ldots, u_N\}$ composta da N unità elementari $u_1, \ldots, u_N$. Queste saranno sempre assunte *identificabili*: ad ognuna di esse può essere assegnata un'*etichetta* che la *identifica univocamente* . Per semplicità, assumiamo che all'unità (elementare) u_i sia associata l'etichetta i. In vista della corrispondenza biunivoca tra unità di osservazione (reali) e etichette, risulta equivalente parlare dell'*unità* u_i e dell'*unità di etichetta* i. Per brevità, e senza perdita di generalità, si identificherà d'ora in avanti ogni unità elementare con la propria etichetta, e si userà la locuzione abbreviata *unità* i in luogo di quella completa *unità di etichetta* i. In forza della corrispondenza dianzi stabilita, l'insieme delle N etichette $I_N = \{1, \ldots, N\}$ verrà d'ora in poi considerato come popolazione di riferimento.

Per ogni unità i sono definite le modalità di uno o più caratteri. Per semplicità, faremo prevalentemente riferimento al caso di un solo carattere, essendo pressoché immediata l'estensione a più caratteri.

Dato un carattere $\mathcal{Y}$, indichiamo con y_i la modalità da esso assunta in corrispondenza dell'unità i. Per il momento, assumeremo che non vi siano errori di misurazione: se l'unità i è inclusa nel campione, la modalità y_i può essere osservata senza errore.

L'*osservazione completa* (o *osservazione etichettata*, o *modalità etichettata*) dell'unità i è la coppia (i, y_i), ossia la coppia costituita dall'unità e dalla corrispondente modalità. Essa conserva l'informazione relativa non solo alla modalità osservata, ma anche all'unità a cui si riferisce, e costituisce il *dato statistico di base*.

Per l'intera popolazione, sono definite le N coppie (osservazioni etichettate) (i, y_i), $i = 1, \ldots, N$. Esse sono equivalenti al vettore (colonna, per convenzione) $\boldsymbol{Y}_N = (y_1 \cdots y_N)^T$, la cui componente i-ma è la modalità y_i dell'unità i. In questo modo, il vettore $\boldsymbol{Y}_N$ è costituito da *modalità etichettate*, contenenti l'informazione relativa non solo alle modalità del carattere $\mathcal{Y}$, ma anche alle unità corrispondenti.

Il vettore $\boldsymbol{Y}_N$ è il *parametro della popolazione*, in quanto individua univocamente il modo in cui il carattere $\mathcal{Y}$ si manifesta nella popolazione. La conoscenza di $\boldsymbol{Y}_N$, ossia l'osservazione (senza errore) delle componenti del vettore $\boldsymbol{Y}_N$ porta ad una perfetta conoscenza delle modalità con cui il carattere $\mathcal{Y}$ si manifesta su tutte le unità della popolazione di interesse.

Nei casi reali, il parametro della popolazione $\boldsymbol{Y}_N$ è in genere incognito. Indicheremo con Ω_N l'insieme di tutti i "valori" che $\boldsymbol{Y}_N$ può assumere. L'insieme Ω_N è lo *spazio dei parametri*. Il caso tipico (anche se tutt'altro che esclusivo) è quello in cui ogni modalità y_i è un numero reale, per cui $\boldsymbol{Y}_N$ è un qualunque punto di $\mathbb{R}^N$: $\Omega_N = \mathbb{R}^N$.

Come anticipato, un *parametro statistico* (di interesse) θ è una funzione delle modalità che costituiscono il vettore $\boldsymbol{Y}_N$. In simboli: $\theta = \theta(\boldsymbol{Y}_N) = \theta(y_1, \ldots, y_N)$. Esempi molto semplici, ma importanti, di parametri statistici sono la *media della popolazione*, per la quale si userà sempre il simbolo μ_y

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.1}$$

e la *varianza della popolazione*, per la quale verrà impiegato il simbolo σ_y^2

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \mu_y^2. \tag{1.2}$$

Esempio 1.1 (Popolazioni dicotomiche). Un caso speciale molto importante è quello in cui si vuole studiare la presenza/assenza di un attributo A sulle unità della popolazione. In questo caso il carattere $\mathcal{Y}$ assume le due sole modalità 1 e 0, indicanti rispettivamente la presenza e l'assenza dell'attributo A. In simboli:

$$y_i = \begin{cases} 1 \ \textit{se l'unità i possiede l'attributo } A \\ 0 \ \textit{altrimenti} \end{cases} \qquad \textit{per ciascuna unità } i = 1, \ldots, N.$$

Lo spazio dei parametri Ω_N è l'insieme $\{0, 1\}^N$ delle N-ple le cui componenti sono uguali a 0 o a 1. Indichiamo con N_A il numero di unità della popolazione che presentano l'attributo A, e sia $P_A = N_A/N$ la proporzione di unità della popolazione che presentano l'attributo A. La media μ_y si riduce alla proporzione ϑ_A:

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_A}{N} = P_A.$$

Tenendo poi conto che $y_i^2 = y_i$, la varianza σ_y^2 è uguale a:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \mu_y^2 = \frac{1}{N}\sum_{i=1}^{N} y_i - P_A^2 = P_A - P_A^2 = P_A(1 - P_A). \qquad \square$$

Nel caso in cui i caratteri di interesse siano due o più, l'approccio è del tutto simile. Per semplicità di notazione, ci limitiamo al caso di due caratteri $\mathcal{X}$, $\mathcal{Y}$. In corrispondenza di un'unità i si ha ora la coppia di modalità (x_i, y_i), assunte rispettivamente da $\mathcal{X}$ e da $\mathcal{Y}$, $i = 1, \ldots, N$.

Le modalità etichettate dell'unità i sono date dalla terna (i, x_i, y_i), costituita dall'unità e dalla corrispondente coppia di modalità dei due caratteri $\mathcal{X}$, $\mathcal{Y}$. Esattamente come nel caso precedente, essa contiene l'informazione relativa sia alle modalità, sia all'unità a cui si riferiscono. Per la

popolazione I_N sono pertanto definite le N terne (osservazioni etichettate) (i, x_i, y_i), $i = 1, \ldots, N$. Esse sono equivalenti ai due vettori (colonna) $\boldsymbol{Y}_N = (y_1 \cdots y_N)'$, $\boldsymbol{X}_N = (x_1 \cdots x_N)'$, o, il che è lo stesso, alla matrice (per convenzione di N righe e due colonne) $(\boldsymbol{X}_N, \boldsymbol{Y}_N)$.

Parametri statistici di interesse sono funzioni delle modalità che costituiscono la matrice $(\boldsymbol{X}_N, \boldsymbol{Y}_N)$. Esempi molto semplici di parametri di interesse, oltre a quelli univariati già introdotti, sono la covarianza tra i due caratteri nella popolazione in esame:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)\,(y_i - \mu_y) = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \mu_x\, \mu_y$$

e il loro coefficiente di correlazione lineare

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\, \sigma_y}$$

in cui la notazione è quella introdotta in (1.1) e (1.2).

1.4 Popolazioni suddivise in sottopopolazioni

Una popolazione I_N può essere suddivisa in M sottoinsiemi distinti di unità, ognuno dei quali costituisce una sua *sottopopolazione*. Indichiamo con

- $I^1_{N_1}$ la prima sottopopolazione, di N_1 unità;
- $I^2_{N_2}$ la seconda sottopopolazione, di N_2 unità;
- $\ldots$ $\ldots$
- $I^M_{N_M}$ la M-ma sottopopolazione, di N_M unità.

Dal punto di vista insiemistico, le M sottopopolazioni $I^g_{N_g}$, $g = 1, \ldots, M$, costituiscono una *partizione* della popolazione I_N: ogni unità di I_N deve necessariamente appartenere ad uno e uno sola delle sottopopolazioni in cui I_N è suddivisa. Formalmente questo significa che l'unione delle varie sottopopolazioni ricostruisce la popolazione totale:

$$I^1_{N_1} \cup I^2_{N_2} \cup \cdots \cup I^M_{N_M} = I_N$$

e che le sottopopolazioni non hanno unità in comune (sono *due a due disgiunte*):

$$I^g_{N_g} \cap I^h_{N_h} = \emptyset \text{ per ciascun } g \neq h;\ g,\ h = 1, \ldots, M.$$

Chiaramente, si deve avere $N_1 + \cdots + N_M = N$. Indichiamo con $w_g = N_g/N$ il *peso* della sottopopolazione g-ma ($g = 1, \ldots, M$). È immediato verificare che valgono le seguenti due relazioni:

$$0 \leqslant w_g \leqslant 1 \text{ per ogni } g = 1, \ldots, M; \quad w_1 + \cdots + w_M = 1.$$

Per comodità di notazione, ciascuna unità della popolazione può essere identificata tramite una doppia etichetta (g, i), in cui:

- g $(= 1, \ldots, M)$ indica l'etichetta che identifica la sottopopolazione;
- i $(= 1, \ldots, N_g)$ identifica l'unità nell'ambito della sottopopolazione a cui appartiene.

Coerentemente con questa simbologia, sia y_{gi} la modalità dell'unità i $(= 1, \ldots, N_g)$ della sottopopolazione g-ma $(g = 1, \ldots, M)$, e siano

$$\mu_{yg} = \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi}\,;\ g = 1, \ldots, M$$

$$\sigma^2_{yg} = \frac{1}{N_g} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2\,;\ g = 1, \ldots, M$$

rispettivamente la media e la varianza del carattere $\mathcal{Y}$ nella sottopopolazione g-ma. La successiva proposizione riassume le proprietà essenziali della media e della varianza di $\mathcal{Y}$ quando la popolazione è suddivisa in sottopopolazioni.

Proposizione 1.1. *Valgono i seguenti due risultati:*

- *la media della popolazione è pari alla media delle medie delle sottopopolazioni (ponderate con i pesi delle stesse):*

$$\mu_y = \sum_{g=1}^{M} w_g\, \mu_{yg}\,; \tag{1.3}$$

- *la varianza della popolazione è uguale alla somma* (*a*) *della media delle varianze delle sottopopolazioni e* (*b*) *della varianza delle medie delle sottopopolazioni (sempre ponderate con i propri pesi):*

$$\sigma^2_y = \sum_{g=1}^{M} w_g\, \sigma^2_{yg} + \sum_{g=1}^{M} w_g\, (\mu_{yg} - \mu_y)^2. \tag{1.4}$$

Dimostrazione. La dimostrazione della (1.3) è immediata:

$$\begin{aligned}
\mu_y &= \frac{1}{N} \sum_{g=1}^{M} \sum_{i=1}^{N_g} y_{gi} \\
&= \sum_{g=1}^{M} \frac{N_g}{N} \left\{ \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi} \right\} \\
&= \sum_{g=1}^{M} w_g\, \mu_{yg}.
\end{aligned}$$

Per quanto riguarda la (1.4), è sufficiente osservare che:

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g}(y_{gi}-\mu_y)^2 \\
&= \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g}\{(y_{gi}-\mu_{yg})+(\mu_{yg}-\mu_y)\}^2 \\
&= \frac{1}{N}\sum_{g=1}^{M}\left\{\sum_{i=1}^{N_g}(y_{gi}-\mu_{yg})^2+\sum_{i=1}^{N_g}(\mu_{yg}-\mu_y)^2\right. \\
&\quad \left.+2\,(\mu_{yg}-\mu_y)\sum_{i=1}^{N_g}(y_{gi}-\mu_{yg})\right\} \\
&= \sum_{g=1}^{M}\frac{N_g}{N}\left\{\frac{1}{N_g}\sum_{i=1}^{N_g}(y_{gi}-\mu_{yg})^2\right\}+\sum_{g=1}^{M}\frac{N_g}{N}(\mu_{yg}-\mu_y)^2 \\
&= \sum_{g=1}^{M} w_g\,\sigma_{yg}^2+\sum_{g=1}^{M} w_g\,(\mu_{yg}-\mu_y)^2. \qquad \square
\end{aligned}$$

1.5 Liste di unità di campionamento

Come già accennato nella Sezione 1.2, l'*accesso* alla popolazione di unità di osservazione avviene tramite una *lista* (*frame*) di *unità di campionamento* (*sampling units*). In generale, una lista è un qualunque meccanismo che permette di accedere alle unità della popolazione e di osservarle. Le singole entità che compongono la lista sono le *unità di campionamento*, in contrapposizione alle unità di osservazione della popolazione, su cui sono definite le modalità del(i) carattere(i) oggetto di interesse. Tramite un'opportuna regola di selezione si sceglie un campione di unità di campionamento, e tramite esse si accede alle corrispondenti unità di osservazione della popolazione.

Esempi molto semplici di liste che permettono l'accesso a popolazioni quali quelle degli individui residenti in Italia, delle aziende, etc. sono di seguito riportati:

- Anagrafe delle famiglie (per comune).
- Liste elettorali (per comune o per sezione elettorale).
- Elenchi degli abbonati alla rete di telefonia fissa (per comune). La loro copertura della popolazione è assai elevata, benché un pò erosa dalla diffusione dei telefoni cellulari. Ad ogni modo, la disponibilità di elenchi su CD favorisce tanto la formazione sistematica di campioni quanto il contatto di individui tramite telefono.

- Lista degli studenti iscritti a scuole pubbliche di tutti gli ordini (per scuola o per facoltà universitaria).
- Albi professionali provinciali previsti per professionisti di varie categorie (medici, avvocati, attuari, commercialisti, etc.).
- Archivi ufficiali quali il registro delle imprese fornito dalla Unione delle Camere di Commercio Italiane. Queste liste sono diventate nel tempo sempre piuttosto obsolete e non molto attendibili, specie a fini di selezione di un campione di aziende. Recentemente si è iniziato a integrarle con liste quali quelle delle Pagine Gialle (comunque abbastanza lacunose), di aziende produttrici di energia elettrica, e con le liste dell'INPS (costruite per fini previdenziali). Sempre per quanto riguarda le imprese che operano sul territorio italiano, la lista più completa è indubbiamente l'archivio ASIA utilizzato dall'ISTAT.
- Liste di unità territoriali amministrative di vario tipo, quali comuni, aziende sanitarie locali, distretti scolastici, distretti elettorali, sezioni di censimento, etc.
- Liste di organi vari, quali scuole, ospedali, agenzie bancarie, enti pubblici, etc.

Il caso più semplice, in linea di principio, è quello in cui nella lista sono elencate le unità di osservazione della popolazione, le quali coincidono con le unità di campionamento. In questo caso la selezione di unità di osservazione può avvenire direttamente dalla lista delle corrispondenti unità di campionamento, per cui si parla di *campionamento diretto* di unità della popolazione. Idealmente, una lista per il campionamento diretto dovrebbe identificare (ad es. tramite una etichetta) *tutte* le unità della popolazione. Inoltre, una volta che un'unità viene scelta, dovrebbe permettere di *localizzarla* e di *contattarla*. Requisiti addizionali importanti sono i seguenti:

- ogni unità della popolazione dovrebbe comparire *una sola volta* nella lista;
- nella lista dovrebbero comparire *solo* le unità della popolazione.

Non sempre i requisiti sopra elencati sono soddisfatti. Consideriamo ad esempio un'azienda produttrice di programmi televisivi, che vuole decidere se lanciare o meno un nuovo canale di tv via cavo specializzato in programmi per giovani e giovanissimi (cartoni animati, programmi di giochi, telefilm, programmi musicali, etc.). L'azienda deve in primo luogo valutare il proprio mercato potenziale, ed in particolare deve avere un'idea di quante famiglie sono disposte ad abbonarsi al canale, e a quale prezzo. La popolazione di riferimento, in questo caso, è quella delle famiglie residenti in Italia ed in cui vi sia almeno un bambino o un ragazzo (tali famiglie sono le nostre unità di osservazione). Un'idea molto semplice potrebbe essere quella di: (a) selezionare un campione di famiglie italiane con almeno un bambino o un ragazzo; (b) accertare la disponibilità delle famiglie del campione a sottoscrivere un abbonamento, e a quali condizioni economiche. Gli elenchi telefonici forniscono una lista per il campionamento diretto, su cui si possono fare le seguenti osservazioni. (i) Non tutte le famiglie che vivono in Italia possiedono il te-

lefono, per cui la nostra lista non contiene tutte le unità della popolazione. Per le famiglie di interesse (quelle con bambini) si tratta tutto sommato di un problema abbastanza marginale, in quanto il numero di famiglie con bambini che non possiedono telefono fisso può essere considerato piccolo. (*ii*) Vi sono famiglie con due o più numeri telefonici, per cui vi sono unità (famiglie) che compaiono due o più volte nella lista. (*iii*) Non in tutte le famiglie dell'elenco telefonico vi è (almeno) un bambino o un ragazzo. Questo significa che la nostra lista contiene (molte) unità che non fanno parte della popolazione di interesse.

La disponibilità di liste per il campionamento diretto di unità (di osservazione) della popolazione non è, tutto sommato, molto frequente nella pratica applicativa. Spesso le unità di campionamento non coincidono con quelle di osservazione, nel senso che possono essere "più grandi" (ogni unità di campionamento è composta da più unità di osservazione) o "più piccole" (ogni unità di osservazione è composta da più unità di campionamento).

Un esempio molto semplice in proposito è quello, già accennato nella Sezione 1.2, della rilevazione ISTAT delle forze di lavoro. La popolazione di riferimento è quella degli individui residenti in Italia, che sono le unità di osservazione. L'accesso a tale popolazione avviene tramite la lista delle famiglie, che sono quindi le unità di campionamento. Più in dettaglio, tramite una opportuna procedura si seleziona un campione di famiglie, e si osservano tutti gli individui appartenenti alle famiglie selezionate. In questo caso ogni unità di campionamento è composta da più unità di osservazione.

Un esempio in cui succede esattamente l'opposto è quello riportato in Särndal *e altri* (1993), che riguarda l'indagine sui redditi delle famiglie svedesi. La popolazione di interesse è quella delle famiglie residenti in Svezia, le quali sono le unità di osservazione. Ora, in Svezia non esiste una lista affidabile delle famiglie residenti, mentre esiste una buona lista degli individui residenti. Pertanto, come unità di campionamento vengono usati gli individui, che permettono di identificare le famiglie a cui appartengono, e di osservare il corrispondente reddito. Chiaramente, in questo caso ogni unità di osservazione è costituita da più unità di campionamento.

Un caso speciale molto importante di unità di campionamento è quello delle *unità areali*. Queste sono di particolare importanza (e utilità) quando le unità di osservazione della popolazione di interesse si trovano in un dato territorio. Per accedere ad esse si può allora pensare di (*i*) suddividere il territorio in aree (che quindi sono le unità di campionamento); (*ii*) selezionare un campione di unità territoriali; (*iii*) osservare (alcune delle o tutte le) unità che vivono nelle aree campione. Questo modo di procedere è particolarmente utile quando non si dispone di una lista delle unità di osservazione, oppure quando liste di unità sono disponibili solo separatamente per area territoriale. Un esempio in proposito è la rilevazione ISTAT delle forze di lavoro, in cui il territorio italiano è suddiviso in province, e ogni provincia in gruppi di comuni. Liste di famiglie sono disponibili separatamente per ogni comune, mentre è molto più difficile (per ragioni di tempo e di costo) avere una lista completa delle

famiglie italiane. Pertanto, risulta conveniente selezionare un campione di comuni, e poi utilizzare le liste comunali per selezionare un campione di famiglie.

A questo punto, siamo in grado di distinguere tra *popolazione obiettivo* e *popolazione da lista*. La *popolazione obiettivo* (*target population*) è l'insieme di tutte le unità (di osservazione) che formano la popolazione di riferimento, sulla quale si vogliono raccogliere informazioni. La *popolazione da lista* (*frame population*) è invece l'insieme di tutte le unità di osservazione a cui si può accedere (e che possono quindi essere osservate) tramite la lista delle unità di campionamento. Il caso ideale è ovviamente quello in cui la popolazione da lista coincide con quella obiettivo. Rispetto alla popolazione obiettivo la lista deve possedere le seguenti caratteristiche:

- completezza: deve contenere tutte le unità della popolazione obiettivo;
- aggiornamento: non deve contenere unità estranee, duplicazioni e ogni unità deve essere distinguibile dalle altre e individuabile.

Quando ciò non accade, si è in presenza di *imperfezioni di lista* (*frame imperfections*; Fig. 1.1). In linea di principio, le principali imperfezioni di lista sono di tre tipi.

- *Sottocopertura.* Si ha quando la popolazione obiettivo contiene unità che non sono nella popolazione da lista. Un qualunque campione non conterrà nessuna di queste unità, che non hanno quindi alcuna possibilità di essere osservate. In altre parole, una parte della popolazione obiettivo viene trascurata, con seri rischi di effetti distorsivi se le sue caratteristiche differiscono dalla parte della popolazione obiettivo su lista.
- *Sovracopertura.* Si ha quando la popolazione da lista contiene unità che non sono nella popolazione obiettivo. Il rischio che si corre in questo caso è di osservare unità che non interessano, con dispendio di tempo e risorse materiali.
- *Duplicazioni.* Si hanno quando una stessa unità compare più volte nella lista. Imperfezioni dovute a duplicazioni sorgono principalmente quando la lista delle unità di campionamento è costruita a partire da due o più

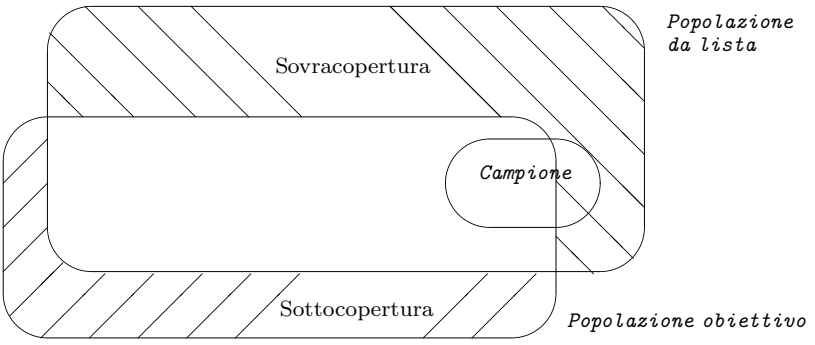


Fig. 1.1 Imperfezioni di lista (sottocopertura e sovracopertura)

"sottoliste", che potrebbero avere unità in comune. Anche di questo fatto bisogna tener conto se si vuole avere una corretta idea di come estendere all'intera popolazione i risultati osservati nel campione.

La coincidenza tra popolazione obiettivo e popolazione da lista non sempre (anzi, abbastanza raramente) ha luogo in pratica. È però di fondamentale importanza che le imperfezioni di lista siano di entità lieve. Gravi imperfezioni di lista (soprattutto per la sottocopertura) possono condurre a conclusioni fuorvianti per quel che riguarda la popolazione obiettivo.

1.6 Rilevazioni statistiche e indagini statistiche

Per evitare futuri equivoci, ci sembra a questo punto opportuno distinguere l'aspetto più esteso di *indagine statistica* da quello più ristretto di *rilevazione statistica*, al quale faremo riferimento d'ora in avanti.

1. Per *rilevazione statistica* (totale o campionaria) intenderemo l'attività che, rivolta alla produzione di dati statistici, percorre le seguenti fasi.
   - *Piano di raccolta dei dati unitari.* Richiede la definizione degli obiettivi dell'indagine, dei mezzi e metodi per attuarla e dei relativi tempi e costi. In questa fase vanno specificati elementi chiave quali la popolazione di riferimento, i caratteri e i parametri di interesse, la lista (o le liste) delle unità di campionamento, le informazioni *a priori* sulla popolazione (inclusi eventualmente alcuni caratteri noti da rilevazioni precedenti), il disegno di campionamento (ossia la regola di selezione delle unità campionarie), le procedure per l'osservazione dei caratteri di interesse (formulazione del questionario, specificazione del tipo di intervista, etc.), gli stimatori da usare.
   - *Messa a punto e controllo del piano di raccolta dati.* Questa fase (se prevista) viene effettuata tramite una *indagine pilota* che consente di mettere a punto il questionario, scoprire eventuali problemi nelle liste di campionamento, nella rilevazione, etc.
   - *Raccolta dei dati.* Viene selezionato il campione (se la rilevazione è campionaria), e sono osservate le modalità delle unità campionate. L'operazione fondamentale svolta in questa fase consiste nella misurazione delle modalità dei caratteri di interesse, in genere tramite somministrazione di un questionario agli intervistati.
   - *Sistemazione e prima elaborazione dei dati.* I dati osservati vengono preparati per analisi successive. Ciò richiede l'esecuzione di una serie di operazioni, quali la codifica dei dati osservati, la loro trascrizione su supporto informatico e controllo logico di coerenza dei dati (*editing*), l'integrazione di dati mancanti (dovuti ad intervistati che non hanno fornito nessuna risposta, o non hanno risposto ad alcune domande del questionario), etc.

   - *Riduzione parametrica dei dati.* In questa fase vengono calcolate le stime dei parametri di interesse della popolazione, così come delle misure di precisione di tali stime. Spesso vengono anche eseguite ulteriori analisi quali il confronto di sottogruppi di unità della popolazione, l'analisi di eventuali relazioni statistiche (regressione, correlazione, connessione e altro) tra caratteri, etc.
2. Per *indagine statistica* intenderemo l'attività più completa che, dopo le fasi di:
   - progettazione di una *ricerca statistica*;
   - acquisizione delle informazioni occorrenti (*stime*) in base ai risultati di una *rilevazione statistica* (pianificata ed eseguita nell'ambito della stessa ricerca o svolta da altre istituzioni);

   procede alle seguenti ulteriori fasi:
   - analisi statistiche di ipotesi (in genere tramite *test* statistici) in merito all'indagine;
   - modellizzazione descrittiva e operativa (costruzione di un modello teorico e suo uso a fini previsivi) per il fenomeno oggetto di indagine.

Mentre un'indagine statistica ha in genere un fine ben preciso, che esige un contenuto spesso molto specifico dei dati da osservare, una rilevazione statistica svolta come rilevazione a sé stante è progettata come *indagine statistica di servizio* (o generale), ovvero destinata a fornire parametri strutturali di base che permettano ai ricercatori di approfondire proprie analisi e modellizzazioni nell'ambito di proprie indagini statistiche. Un esempio in proposito è fornito dalla rilevazione delle forze di lavoro. Essa fornisce da un lato parametri (tasso di disoccupazione, ammontare e composizione delle forze di lavoro, etc.) utili per studiare l'andamento del mercato del lavoro e per valutare l'effetto di politiche economiche. Dall'altro, i dati forniti dalla rilevazione delle forze di lavoro sono ampiamente usati da ricercatori per verificare teorie economiche sul mercato del lavoro, per analizzare e prevedere (assieme eventualmente ad altri dati) l'andamento della domanda e dell'offerta di lavoro, etc.

1.7 Fonti di errore e distorsioni

Nelle operazioni effettuate in ogni rilevazione statistica, e soprattutto in quelle su larga scala, vi è la possibilità di errori e distorsioni che possono grandemente influenzare i risultati che si ottengono. Qui di seguito viene effettuata una breve discussione delle principali categorie di tali errori e distorsioni.

1. *Distorsioni ed errori nel campione.* In questa categoria sono inclusi le distorsioni e gli errori che dipendono da imprecisioni di lista (già discusse in precedenza) e dal processo di selezione delle unità del campione.
   - La più grave fonte di distorsione dovuta a imprecisioni di lista è quella dovuta a *sottocopertura*: alcune delle unità della popolazione obiettivo non compaiono nella popolazione da lista, e quindi *non possono esse-*

re selezionate nel campione. Chiaramente, questo potrebbe essere un difetto potenzialmente molto grave, che potrebbe produrre distorsioni dovute al trascurare sistematicamente una parte della popolazione. Inoltre, anche se trascurabile a livello di intera popolazione, la sottocopertura potrebbe essere rilevante quando si esaminano singole sottopopolazioni. Un'altra fonte di possibili distorsioni, di cui bisogna tener conto in fase di costruzione di stime, è la presenza di *duplicazioni.* Questo accade soprattutto quando la lista di campionamento deriva dalla fusione di più liste aventi unità in comune. Per quanto riguarda la *sovracopertura* della lista, invece, questa non ha di per sé effetti distorsivi. Tuttavia, può portare a contattare unità che non interessano la rilevazione, con un conseguente dispendio di risorse che potrebbero essere utilmente impiegate.

- Gli *errori campionari* sono tutti gli errori dovuti all'osservare soltanto una parte della popolazione. Questo tipo di errore è ineliminabile nelle rilevazioni campionarie, ma può essere grandemente ridotto attraverso la scelta della numerosità campionaria, della regola di selezione delle unità che costituiscono il campione e dello stimatore.

2. *Distorsioni ed errori nel processo di acquisizione dei dati.* Le principali fonti di errori, a questo livello, sono gli errori di misurazione e le mancate risposte.
   - Gli *errori di misurazione* consistono nel fatto che non si osservano esattamente le modalità dei caratteri oggetto di interesse, in quanto le osservazioni contengono errori di misura. Questi sono dovuti a svariate ragioni. Anzitutto, l'intervistato potrebbe (intenzionalmente o meno) fornire risposte errate; questo può accadere, ad es., in indagini sul reddito, o su argomenti "delicati", quali consumi di droga o alcool, etc. In secondo luogo, il questionario potrebbe essere poco chiaro, o potrebbe essere l'intervistatore a formulare la domanda in modo erroneo, o trascrivere in maniera errata la risposta. Ancora, l'intervistatore potrebbe influenzare l'intervistato in modo da distorcerne la risposta. In ogni caso, gli errori di misurazione sono una fonte molto seria di errore, e potrebbero avere effetti molto seri sulla qualità delle stime dei parametri della popolazione.
   - Le *mancate risposte* si hanno quando l'intervistato è irreperibile, oppure non vuole rispondere o non è in grado di rispondere ad alcune o a tutte le domande del questionario. Le distorsioni dovute a mancate risposte sono particolarmente gravi, e possono inficiare grandemente i risultati di un'indagine. Di esse bisogna tener conto con grande attenzione nella fase di costruzione di stime dei parametri.
3. *Errori nell'elaborazione dei dati.* Si tratta degli errori dovuti alle operazioni di codifica e trascrizione dei dati su supporto informatico, di errori non rilevati nella fase di *editing*, nell'integrazione dei dati mancanti, etc.

1.8 Come non progettare una rilevazione campionaria

Gli errori a cui è soggetta una rilevazione vanno sempre tenuti ben presenti, perchè, come sottolineato più volte, possono addirittura inficiarne i risultati. Un esempio celebre è quello dell'indagine promossa dal *Literary Digest*. Nelle elezioni presidenziali statunitensi del 1936 molti giornali organizzarono un sondaggio presso il corpo elettorale della popolazione finalizzato alla previsione dei risultati delle elezioni. La sfida avvenne tra il candidato repubblicano Alfred M. Landon e quello democratico Franklin. D. Roosevelt. La nota rivista *Literary Digest*, che aveva correttamente previsto i risultati delle quattro elezioni presidenziali americane precedenti (1920, 1924, 1928, 1932) ricorrendo a sondaggi d'opinione, previde che Alfred M. Landon avrebbe ottenuto il 55% dei voti contro il 41% del presidente in carica, Franklin D. Roosevelt.

Nella indagine condotta dal Literary Digest 10 milioni di *fac-simile* di schede elettorali furono inviate per posta a nominativi estratti dagli elenchi telefonici e dai registri automobilistici. Dei 10 milioni circa 2.4 milioni risposero al sondaggio. Malgrado l'enormità della numerosità del campione l'esito delle elezioni smentì completamente il pronostico. Franklin D. Roosevelt divenne presidente con il 61% delle preferenze contro il 37% del candidato repubblicano. Il clamoroso errore distrusse la credibilità della rivista che cessò la pubblicazione nel 1938.

L'errore nella previsione commesso dal Literary Digest è essenzialmente riconducibile a due cause principali:

- *Distorsione dovuta alla selezione del campione (errore di sottocopertura).* Il Literary Digest aveva compilato la lista della popolazione utilizzata per l'estrazione del campione sfruttando gli elenchi degli abbonati telefonici e dei proprietari di automobili. Tali elenchi sovrarappresentavano i ceti più abbienti, che tendevano a votare prevalentemente repubblicano, e sottorappresentavano la popolazione dei votanti del partito democratico. In termini formali, la popolazione da lista (elettori che disponevano di telefono e di automobile) differiva sostanzialmente dalla popolazione obiettivo (tutti gli elettori).
- *Distorsione dovuta alle mancate risposte.* Il basso tasso di risposta combinato con la distorsione dovuta ai non rispondenti aveva completamente falsato i risultati della rilevazione. Un tasso di risposta del 24% è troppo basso per ottenere risultati attendibili dei parametri di interesse della popolazione a meno che non sia possibile assumere che i 7.6 milioni di non rispondenti abbiano la stessa opinione dei rispondenti. Nella pratica delle indagini campionarie, non è lecito in generale assumere che i rispondenti siano simili ai non rispondenti.

L'iniziale distorsione presente nel campione è stata accentuata dal fatto che le persone appartenenti ai ceti più abbienti e che tendevano ad essere sostenitori di Landon erano anche più propensi a rispondere al sondaggio. D'altra parte

anche se tutti gli elettori appartenenti al campione avessero risposto, non si sarebbe annullata la distorsione dovuta alla sottocopertura.

Contemporaneamente al sondaggio della Literary Digest, George Gallup, utilizzando un campione di poche migliaia di americani, predisse correttamente la vittoria di Roosevelt. Quindi un campione grande non sempre fornisce risultati più attendibili di un campione di dimensione più esigua. Non è importante solo la dimensione del campione, ma ancor più la sua composizione. In conclusione, nel caso del Literary Digest il campione, pur numericamente enorme, non si rivelò rappresentativo della popolazione a causa dell'inadeguatezza delle liste utilizzate per la sua estrazione. A tale effetto, si combina l'effetto distorsivo dovuto al fenomeno dell'autoselezione dei rispondenti.

1.9 Campionamento non probabilistico

La distinzione principale che occorre effettuare sul concetto di campione è quella tra campione probabilistico e non probabilistico.

Si parla di campione probabilistico quando la selezione del campione avviene sulla base di una legge di probabilità (disegno campionario) nota a priori perché prefissata dallo statistico in fase di progettazione della rilevazione. Per poter effettuare un campionamento probabilistico è fondamentale disporre della lista di unità che costituiscono la popolazione oggetto di studio. Le unità della popolazione sono selezionate dalla lista secondo un meccanismo casuale, e ogni unità della popolazione ha una probabilità nota e non nulla di essere inclusa nel campione.

Sono non probabilistici i campioni che non soddisfano la precedente condizione. Nel campionamento non probabilistico la scelta delle unità campionarie viene effettuata sulla base di criteri di comodo e di praticità e/o sulla base di informazioni a priori relative alle caratteristiche della popolazione di interesse. Nonostante questi metodi non escludano la possibilità di ottenere stime accurate delle grandezze di interesse della popolazione (medie, proporzioni, etc.), è impossibile valutare la precisione delle stime. Il campionamento non probabilistico non consente di valutare l'accuratezza dei risultati ottenuti a livello campionario, e le loro (eventuali) relazioni con le corrispondenti grandezze a livello di popolazione.

Supponiamo ad esempio di voler effettuare un sondaggio per valutare la qualità del servizio di mensa di una scuola. A questo scopo si decide di intervistare i primi 100 studenti che si presentano alla mensa. Chiaramente il campione selezionato non ha nulla di casuale essendo costruito in modo accidentale. Nel campionamento probabilistico il concetto di casualità è strettamente connesso a quello di probabilità: selezionare le unità a caso non vuol dire selezionarle "a casaccio", ma bensì selezionarle secondo una procedura predefinita e casualizzata in modo controllata dallo statistico.

Nel campionamento probabilistico la condizione di casualità è una condizione necessaria per poter ricondurre alla popolazione, attraverso la teo-

ria della probabilità, i risultati ottenuti dal campione con un certo grado di affidabilità.

Il campionamento non probabilistico è utilizzato nelle indagini in cui non è disponibile una lista di unità da cui estrarre il campione o il costo di costruzione di tale lista è proibitivo, e nelle indagini in cui si vuole contenere il costo di raccolta delle informazioni. Metodi di campionamento non probabilistici includono (tra gli altri) il *campionamento ragionato*, il *campionamento per quote*, e il *campionamento a valanga*.

Nel *campionamento ragionato* la scelta delle unità da includere nel campione è affidata al giudizio di un esperto. Ad esempio, con riferimento ad una regione italiana supponiamo di voler stimare una determinata caratteristica e che un esperto scelga tre città della regione da cui collezionare i dati su cui basare la stima. L'idea alla base della scelta dell'esperto è che nelle tre città si riscontrino comportamenti analoghi a quelli dell'intera popolazione, così che possano in buona misura "rappresentarla". Chiaramente, poiché la scelta delle unità campionarie non si basa su criteri di casualità bensì sulla competenza dell'esperto, il metodo manca di oggettività. La rappresentatività del campione selezionato dipende fortemente dal livello di conoscenza che l'esperto ha della popolazione oggetto di studio.

Il *campionamento per quote* è frequentemente utilizzato nelle indagini di mercato e nei sondaggi di opinione a causa dei tempi rapidi di realizzazione e dei costi ridotti. Nel campionamento per quote la popolazione viene suddivisa in gruppi omogenei sulla base di variabili strutturali legate alla variabile di interesse (ad esempio: sesso, età, area geografica, etc.). 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. Lo scopo è riprodurre nel campione (relativamente ai gruppi formati) la struttura della popolazione. Si perviene in questo modo alla definizione delle quote, cioé del numero di interviste che ciascun intervistatore deve effettuare in ciascuna classe.

La caratteristica fondamentale del campionamento per quote è che la scelta delle persone da intervistare è completamente demandata all'intervistatore. Chiaramente, la soggettività del criterio di selezione delle unità campionarie da parte dell'intervistatore va a svantaggio della rappresentatività del campione. Per esempio, l'intervistatore potrebbe scegliere di intervistare le persone appartenenti a determinate zone della città per lui più facilmente raggiungibili, le persone più disponibili, le persone appartenenti alla cerchia dei propri conoscenti, evitando di selezionare gli abitanti dei quartieri periferici lontani dalla propria residenza e/o gli abitanti ai piani superiori delle abitazioni.

A volte si cerca di limitare l'arbitrarietà dell'intervistatore introducendo dei vincoli nella scelta delle unità da intervistare, quali ad esempio l'obbligo di compiere prestabiliti itinerari, il divieto di inserire nel campione più unità (individui) facenti parte di uno stesso nucleo abitativo, etc.

Vale la pena sottolineare che nel campionamento per quote gli effetti provocati dalle mancate risposte delle unità contattate dall'intervistatore non

sono controllabili, poiché l'intervistatore completerà sempre il numero di interviste assegnategli contattando nuove persone. Quindi se da una parte è possibile eliminare facilmente le mancate risposte dall'altra si ha l'illusione di eliminarne gli effetti distorsivi sulle stime dei parametri di interesse. Si osservi che le persone che accettano di rispondere potrebbero differire da quelle che non rispondono, con la conseguente introduzione di seri effetti distorsivi.

Il *campionamento a valanga* o a *palla di neve* è utilizzato soprattutto nelle indagini sociologiche che affrontano temi sensibili (omosessualiatà, consumo di droga o alcool, etc.), o nelle indagni su popolazioni rare i cui componenti sono in gran parte ignoti e di difficile reperibilità (clandestini, senzatetto, etc.). Tale campionamento consiste nello scegliere un gruppo iniziale di persone, dalle quali poi risalire ad altre unità appartenenti alla stessa popolazione. Ad esempio, in un'indagine sugli immigrati, si contattano alcuni immigrati clandestini, e poi a fine intervista si chiede loro di indicare i nomi di altri clandestini di loro conoscenza.

Come sottolineato in precedenza, la condizione che ogni unità della popolazione sia caratterizzata da una probabilità nota e non nulla di essere inclusa nel campione riveste un ruolo fondamentale nell'approccio al campionamento probabilistico, ma nella pratica delle indagini campionarie possono esistere delle ragioni che non ne consentono l'applicabilità. Tali metodi assumono, nell'ambito del campionamento da popolazioni finite, una posizione intermedia tra il campionamento probabilistico e il campionamento non probabilistico. Tra questi metodi ricordiamo il *campionamento cut-off*, in cui alcuni elementi della popolazione di interesse sono deliberatamente esclusi dalla selezione campionaria. Chiaramente, il ricorso a tale procedura, che può condurre a distorsioni anche molto gravi dei risultati, è giustificato solo nelle seguenti condizioni:

1. costi eccessivi dovuti alla costruzione di una lista di campionamento per l'intera popolazione in relazione al piccolo guadagno di efficienza che si può ottenere;
2. gli effetti distorsivi sui risultati possono considerarsi trascurabili.

Il campionamento *cut-off* è generalmente utilizzato quando la distribuzione della variabile di interesse nella popolazione è fortemente asimmetrica e non esiste una lista di campionamento affidabile per le "piccole unità. Tali popolazioni sono tipiche delle indagini sulle imprese, in cui una proporzione elevata della popolazione è costituita da piccole imprese (caratterizzate da pochi dipendenti) il cui contributo al valore della variabile di interesse (ad esempio il fatturato) è modesto, e poche grandi imprese. In tali casi si può decidere di escludere dalla selezione del campione le piccole imprese. Si osservi che tale procedura è sconsigliata quando si ha la possibilità di costruire una lista di campionamento della popolazione ad un costo non eccessivo.

2

Campionamento probabilistico

2.1 Disegni campionari: definizione e proprietà di base

In questo e nei successivi paragrafi ci porremo nelle condizioni ideali in cui vi sia perfetta coincidenza tra popolazione obiettivo e popolazione da lista. La notazione che useremo sarà quella del Capitolo 1, in cui si identificano le unità con le loro etichette. Come già detto, se il parametro della popolazione $\boldsymbol{Y}_N$ fosse noto, si potrebbe calcolare il valore che assume un qualunque parametro statistico di interesse. L'osservazione di tutte le modalità $y_1, \ldots, y_N$, ossia l'esecuzione di un censimento, è possibile solo in pochi casi eccezionali. La regola è invece quella delle rilevazioni campionarie, in cui si segue uno schema di base di questo tipo:

1. si seleziona un sottoinsieme di unità della popolazione;
2. si osservano le modalità delle unità in 1;
3. si usano le osservazioni in 2 per cercare di ottenere una qualche "ragionevole approssimazione" di uno o più parametri statistici di interesse.

Definizione 2.1. *Un campione* $\mathbf{s}$ *è un qualunque insieme di unità della popolazione* I_N*. Lo spazio dei campioni* $\mathcal{S}$ *è l'insieme di tutti i campioni che si considerano.*

L'ingrediente decisivo, ovviamente, è il meccanismo, la regola, di selezione del campione. Le più importante regole di selezione di unità della popolazione sono quelle di tipo *probabilistico*, in cui chi progetta la rilevazione ("lo statistico") fissa uno schema probabilistico di selezione delle unità. D'ora in avanti useremo il termine *disegno campionario*, senza altre specificazioni, per indicare proprio schemi di selezione di unità di tipo probabilistico prefissati dallo statistico.

Definizione 2.2. *Un disegno campionario (probabilistico) è una coppia* $(\mathcal{S}, p(\cdot))$ *in cui* $\mathcal{S}$ *è uno spazio dei campioni, e* $p(\cdot)$ *è una distribuzione di*

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_2,

probabilità su $\mathcal{S}$, la quale soddisfa le condizioni:

$$0 < p(\mathbf{s}) \leqslant 1 \text{ per ogni } \mathbf{s} \in \mathcal{S}; \quad \sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) = 1.$$

Un disegno campionario è:

- *con ripetizione* se in almeno un campione $\mathbf{s}$ in $\mathcal{S}$ una stessa unità compare più di una volta;
- *ordinato* se vi sono almeno due campioni $\mathbf{s}_1$, $\mathbf{s}_2$ di $\mathcal{S}$ formati dalle stesse unità, ma poste in ordine differente.

Una distinzione importante nell'ambito dei disegni campionari è quella tra disegni campionari informativi e non informativi. In questo volume saranno considerati unicamente disegni campionari non informativi. Ciò significa che la probabilità con cui il campione viene estratto $p(\mathbf{s})$ non dipende dai valori della variabile di interesse $\mathcal{Y}$ associati alle unità della popolazione.

Per distinguere insiemi ordinati e non ordinati, d'ora in avanti adotteremo una semplice convenzione. Useremo le parentesi tonde per indicare insiemi ordinati (ovvero sequenze), e le parentesi graffe per indicare insiemi non ordinati. Ad es., $\{1, 2\}$ indica l'insieme non ordinato formato dalle unità 1, 2, che quindi è perfettamente equivalente all'insieme $\{2, 1\}$. Con il simbolo $(1, 2)$ indicheremo invece l'insieme ordinato (coppia) formato dalle unità 1, 2. Esso è differente da $(2, 1)$, in quanto i suoi elementi sono posti in ordine diverso.

Esempio 2.1. Si consideri una popolazione di $N = 7$ unità: $I_7 = \{1, 2, \ldots, 7\}$. Supponiamo poi che lo spazio dei campioni sia formato dagli otto campioni:

$$\mathbf{s}_1 = (1, 2, 3),\ \mathbf{s}_2 = (1, 2, 4),\ \mathbf{s}_3 = (5, 6),\ \mathbf{s}_4 = (7),$$
$$\mathbf{s}_5 = (6, 5),\ \mathbf{s}_6 = (1, 2, 2, 3),\ \mathbf{s}_7 = (3, 1, 2),\ \mathbf{s}_8 = (3, 1, 1, 2)$$

con le seguenti probabilità

$$p(\mathbf{s}_1) = 0.1,\ p(\mathbf{s}_2) = 0.15,\ p(\mathbf{s}_3) = 0.15,\ p(\mathbf{s}_4) = 0.05,$$
$$p(\mathbf{s}_5) = 0.2,\ p(\mathbf{s}_6) = 0.05,\ p(\mathbf{s}_7) = 0.1,\ p(\mathbf{s}_8) = 0.2.$$

Si tratta di un disegno campionario ordinato (in quanto, ad es., i due campioni $\mathbf{s}_1$, $\mathbf{s}_7$ sono formati dalle stesse unità ma poste in ordine diverso) e con ripetizioni (perché ad es. nel campione $\mathbf{s}_8$ l'unità 1 compare due volte). □

Esempio 2.2. Consideriamo una popolazione di $N = 7$ unità: $I_7 = \{1, 2, \ldots, 7\}$. Lo spazio dei campioni è formato dai sei campioni:

$$\mathbf{s}_1 = \{1, 2\},\ \mathbf{s}_2 = \{1, 3\},\ \mathbf{s}_3 = \{4\},\ \mathbf{s}_4 = \{2, 3, 5\},\ \mathbf{s}_5 = \{6, 7\},\ \mathbf{s}_6 = \{4, 5, 7\}$$

con le seguenti probabilità

$$p(\mathbf{s}_1) = 0.15,\ p(\mathbf{s}_2) = 0.2,\ p(\mathbf{s}_3) = 0.1,\ p(\mathbf{s}_4) = 0.1,\ p(\mathbf{s}_5) = 0.15,\ p(\mathbf{s}_6) = 0.3.$$

Si tratta di un disegno campionario non ordinato e senza ripetizioni. □

La *numerosità campionaria* $n(\mathbf{s})$ di un campione $\mathbf{s}$ è pari al numero delle unità (non necessariamente distinte) che formano il campione $\mathbf{s}$. La *numerosità campionaria effettiva* $\nu(\mathbf{s})$ di un campione $\mathbf{s}$ è invece il numero di unità *distinte* che formano il campione $\mathbf{s}$. Chiaramente, si ha sempre $\nu(\mathbf{s}) \leqslant n(\mathbf{s})$. Inoltre, è $\nu(\mathbf{s}) = n(\mathbf{s})$ se e solo se nel campione $\mathbf{s}$ non vi sono ripetizioni.

Esempio 2.3. Si consideri l'Esempio 2.1. Il campione $\mathbf{s}_6$ ha una numerosità uguale a 4, ma una numerosità effettiva pari a 3. In simboli: $n(\mathbf{s}_6) = 4$, $\nu(\mathbf{s}_6) = 3$. □

L'*ampiezza media* di un disegno campionario, indicata con $\overline{n}$, è il numero medio di unità contenute nei campioni. In simboli:

$$\overline{n} = E[n(\mathbf{s})] = \sum_{\mathbf{s}\in\mathcal{S}} n(\mathbf{s})\, p(\mathbf{s}).$$

Similmente, l'*ampiezza media effettiva* di un disegno campionario, indicata con $\overline{\nu}$, è il numero medio di unità *distinte* contenute nei campioni:

$$\overline{\nu} = E[\nu(\mathbf{s})] = \sum_{\mathbf{s}\in\mathcal{S}} \nu(\mathbf{s})\, p(\mathbf{s}).$$

Chiaramente, è sempre $\overline{\nu} \leqslant \overline{n}$. Inoltre, l'uguaglianza $\overline{\nu} = \overline{n}$ vale se e solo se il disegno campionario è senza ripetizioni.

Esempio 2.4. Si consideri una popolazione di $N = 5$ unità: $I_5 = \{1, 2, 3, 4, 5\}$, e si supponga che lo spazio dei campioni sia formato dai sette campioni:

$$\mathbf{s}_1 = (1, 2, 1),\ \mathbf{s}_2 = (1, 1, 2, 2),\ \mathbf{s}_3 = (1, 4),\ \mathbf{s}_4 = (4, 5, 3),$$
$$\mathbf{s}_5 = (3, 4, 1, 1),\ \mathbf{s}_6 = (3, 4, 5, 4, 3),\ \mathbf{s}_7 = (4, 1)$$

con le seguenti probabilità

$$p(\mathbf{s}_1) = 0.1,\ p(\mathbf{s}_2) = 0.3,\ p(\mathbf{s}_3) = 0.1,\ p(\mathbf{s}_4) = 0.2,$$
$$p(\mathbf{s}_5) = 0.2,\ p(\mathbf{s}_6) = 0.05,\ p(\mathbf{s}_7) = 0.05.$$

Si tratta di un disegno campionario ordinato e con ripetizioni. Le numerosità dei campioni, e le corrispondenti numerosità effettive sono qui di seguito riportate:

$$n(\mathbf{s}_1) = 3,\ n(\mathbf{s}_2) = 4,\ n(\mathbf{s}_3) = 2,\ n(\mathbf{s}_4) = 3,\ n(\mathbf{s}_5) = 4,\ n(\mathbf{s}_6) = 5,\ n(\mathbf{s}_7) = 2;$$
$$\nu(\mathbf{s}_1) = 2,\ \nu(\mathbf{s}_2) = 2,\ \nu(\mathbf{s}_3) = 2,\ \nu(\mathbf{s}_4) = 3,\ \nu(\mathbf{s}_5) = 3,\ \nu(\mathbf{s}_6) = 3,\ \nu(\mathbf{s}_7) = 2.$$

L'ampiezza media e l'ampiezza media effettiva sono rispettivamente eguali a:

$$\begin{aligned}\overline{n} &= 3\,0.1 + 4\,0.3 + 2\,0.1 + 3\,0.2 + 4\,0.2 + 5\,0.05 + 2\,0.05\\ &= 3.45;\\ \overline{\nu} &= 2\,0.1 + 2\,0.3 + 2\,0.1 + 3\,0.2 + 3\,0.2 + 3\,0.05 + 2\,0.05\\ &= 2.45.\end{aligned}$$

□

La *riduzione* $r(\mathbf{s})$ di un campione $\mathbf{s}$ è l'insieme (non ordinato) delle sue unità distinte. Poiché le unità che compongono la riduzione $r(\mathbf{s})$ sono le unità distinte di $\mathbf{s}$, è immediato concludere che la numerosità di $r(\mathbf{s})$ è null'altro che la numerosità effettiva di $\mathbf{s}$: $n(r(\mathbf{s})) = \nu(\mathbf{s})$.

Dato un disegno campionario $(\mathcal{S}, p(\cdot))$, supponiamo di far corrispondere ad ogni campione "originario" $\mathbf{s} \in \mathcal{S}$ la sua riduzione $r(\mathbf{s})$. Ciò che si ottiene è un nuovo disegno campionario $(\mathcal{S}^*, p^*(\cdot))$, la *riduzione di* $(\mathcal{S}, p(\cdot))$, in cui:

– lo spazio dei campioni $\mathcal{S}^*$ è l'insieme di tutte le riduzioni dei campioni di $\mathcal{S}$:

$$\mathcal{S}^* = \{\mathbf{s}^* = r(\mathbf{s});\ \mathbf{s} \in \mathcal{S}\};$$

– ogni campione $\mathbf{s}^*$ di $\mathcal{S}^*$ ha probabilità pari alla somma delle probabilità dei campioni "originali" $\mathbf{s}$ di $\mathcal{S}$ la cui riduzione è $\mathbf{s}^*$. In simboli, posto $C(\mathbf{s}^*) = \{\mathbf{s} \in \mathcal{S} :\ r(\mathbf{s}) = \mathbf{s}^*\}$, si ha:

$$p^*(\mathbf{s}^*) = \sum_{\mathbf{s} \in C(\mathbf{s}^*)} p(s).$$

Esempio 2.5. Consideriamo il disegno campionario dell'Esempio 2.1, e costruiamo la sua riduzione. Lo spazio dei campioni ridotto $\mathcal{S}^*$ è formato dai campioni:

$$\mathbf{s}_1^* = \{1, 2, 3\},\ \mathbf{s}_2^* = \{1, 2, 4\},\ \mathbf{s}_3^* = \{5, 6\},\ \mathbf{s}_4^* = \{7\}.$$

Essendo poi

$$C(\mathbf{s}_1^*) = \{\mathbf{s}_1, \mathbf{s}_6, \mathbf{s}_7, \mathbf{s}_8\},\ C(\mathbf{s}_2^*) = \{\mathbf{s}_2\},\ C(\mathbf{s}_3^*) = \{\mathbf{s}_3, \mathbf{s}_5\},\ C(\mathbf{s}_4^*) = \{\mathbf{s}_4\}$$

si può anche scrivere:

$$\begin{aligned}
&p^*(\mathbf{s}_1^*) = p(\mathbf{s}_1) + p(\mathbf{s}_6) + p(\mathbf{s}_7) + p(\mathbf{s}_8) = 0.1 + 0.05 + 0.1 + 0.2 = 0.45,\\
&p^*(\mathbf{s}_2^*) = p(\mathbf{s}_2) = 0.15,\\
&p^*(\mathbf{s}_3^*) = p(\mathbf{s}_3) + p(\mathbf{s}_5) = 0.15 + 0.2 = 0.35,\\
&p(\mathbf{s}_4^*) = p(\mathbf{s}_4) = 0.05.
\end{aligned}$$

□

Un disegno campionario è ad *ampiezza costante* se tutti i campioni hanno la stessa numerosità n. In simboli: $n(\mathbf{s}) = n$ per ogni $\mathbf{s} \in \mathcal{S}$. Un disegno campionario è invece ad *ampiezza effettiva costante* se tutti i campioni hanno la stessa numerosità effettiva ν. In simboli: $\nu(\mathbf{s}) = \nu$ per ogni $\mathbf{s} \in \mathcal{S}$.

Esempio 2.6. Sia $I_6 = \{1, \ldots, 6\}$ una popolazione finita di $N = 6$ unità. Il disegno campionario in cui $\mathcal{S}$ è formato dai campioni

$$\mathbf{s}_1 = (1, 2, 2),\ \mathbf{s}_2 = (1, 3, 2),\ \mathbf{s}_3 = (1, 4, 6),\ \mathbf{s}_4 = (4, 5, 3)$$

con le seguenti probabilità

$$p(\mathbf{s}_1) = 0.2,\ p(\mathbf{s}_2) = 0.3,\ p(\mathbf{s}_3) = 0.25,\ p(\mathbf{s}_4) = 0.25$$

è ad ampiezza contante $n = 3$, in quanto tutti i campioni sono formati da tre unità. Esso non è però ad ampiezza effettiva costante, poiché ad esempio è $\nu(\mathbf{s}_1) = 2$ e $\nu(\mathbf{s}_4) = 3$.

Invece, il disegno campionario in cui i campioni che formano $\mathcal{S}$ sono

$$\mathbf{s}_1 = (1, 3),\ \mathbf{s}_2 = (2, 6, 6),\ \mathbf{s}_3 = (5, 4, 5),\ \mathbf{s}_4 = (5, 5, 1)$$

con probabilità

$$p(\mathbf{s}_1) = 0.25,\ p(\mathbf{s}_2) = 0.35,\ p(\mathbf{s}_3) = 0.25,\ p(\mathbf{s}_4) = 0.15,$$

è ad ampiezza effettiva costante $\nu = 2$, poiché tutti i campioni sono formati da due unità distinte. Esso non è ad ampiezza costante, essendo $n(\mathbf{s}_1) = 2$ e $n(\mathbf{s}_2) = 3$. □

2.2 Implementazione di disegni campionari mediante schemi: brevi cenni

Se definire in maniera teorica un disegno campionario è semplice, la selezione effettiva, nella pratica applicativa, di un campione in base ad un dato disegno campionario non sempre è agevole (anzi, lo è piuttosto di rado).

In linea di principio, se il numero di campioni è "piccolo", si può pensare di enumerare i campioni e di sceglierli mediante generazione di un numero pseudo-casuale. Precisamente, si supponga che lo spazio dei campioni sia composto da k campioni: $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$. In primo luogo, i campioni vanno elencati nel modo seguente.

Campione	*Probabilità*	*Probabilità cumulate*
$\mathbf{s}_1$	$p(\mathbf{s}_1)$	$P_1 = p(\mathbf{s}_1)$
$\mathbf{s}_2$	$p(\mathbf{s}_2)$	$P_2 = p(\mathbf{s}_1) + p(\mathbf{s}_2)$
$\mathbf{s}_3$	$p(\mathbf{s}_3)$	$P_3 = p(\mathbf{s}_1) + p(\mathbf{s}_2) + p(\mathbf{s}_3)$
$\cdots$	$\cdots$	$\cdots$
$\mathbf{s}_k$	$p(\mathbf{s}_k)$	$P_k = p(\mathbf{s}_1) + p(\mathbf{s}_2) + \cdots + p(\mathbf{s}_k) = 1$

Si genera poi un numero casuale U, con distribuzione uniforme nell'intervallo $[0, 1]$, e, posto $P_0 = 0$, si procede secondo lo schema riportato qui sotto:

- se $P_0 \leqslant U \leqslant P_1$ si seleziona $\mathbf{s}_1$;
- se $P_1 < U \leqslant P_2$ si seleziona $\mathbf{s}_2$;
- se $P_2 < U \leqslant P_3$ si seleziona $\mathbf{s}_3$;
- $\cdots$
- se $P_{k-1} < U \leqslant P_k$ si seleziona $\mathbf{s}_k$.

Poiché il campione $\mathbf{s}_j$, $j = 1, \ldots, k$ è scelto se e solo se $P_{j-1} < U \leqslant P_j$, la probabilità di selezionare $\mathbf{s}_j$ è pari a:

$$\begin{aligned} Pr(\textit{Selezionare } \mathbf{s}_j) &= Pr(P_{j-1} < U \leqslant P_j) \\ &= P_j - P_{j-1} \\ &= p(\mathbf{s}_1) + \cdots + p(\mathbf{s}_j) - (p(\mathbf{s}_1) + \cdots + p(\mathbf{s}_{j-1})) \\ &= p(\mathbf{s}_j) \end{aligned}$$

qualunque sia $j = 1, \ldots, k$.

Il metodo sopra esposto è molto semplice, ma ha un sostanziale difetto: richiede l'enumerazione dei campioni dello spazio $\mathcal{S}$. Ora, se il numero dei possibili campioni $\mathcal{S}$ è "piccolo", elencarli tutti non presenta particolari difficoltà. Spesso, però, l'elencare tutti i possibili campioni può essere un'operazione improba, o impossibile. Ad esempio, si consideri una popolazione di $N = 1000$ studenti e si supponga di voler selezionare un campione di $n = 100$ di essi, in base ad un disegno (il *disegno semplice*, come si vedrà nel prossimo capitolo) che attribuisce ad ogni sottoinsieme di 100 dei 1000 studenti la stessa probabilità di essere selezionato. Il numero di possibili campioni è $\binom{1000}{100}$. Usando la formula di Stirling ($\log n! \approx n \log n - n + \frac{1}{2} \log n + \frac{1}{2} \log 2\pi$), si vede che

$$\binom{1000}{100} \approx e^{321} > 10000000000000000000000000000000000000000000000000\cdots$$

un numero troppo grande anche per essere solo scritto. Questo significa che in molti casi pratici, e specialmente in quelli che coinvolgono popolazioni numerose, il metodo di selezione sopra esposto, e basato sul preventivo elenco di tutti i possibili campioni, è improponibile.

Nella gran parte dei casi, per selezionare un campione vengono usate procedure che prendono il nome di *schemi campionari*, o *algoritmi di campionamento*. Si tratta di metodi sostanzialmente *ad hoc*, che cercano di selezionare un campione da un dato disegno *senza enumerare tutti i campioni dell'insieme* $\mathcal{S}$, in quanto quest'operazione, nella stragrande maggioranza dei casi concreti, è troppo onerosa sul piano computazionale. In questo caso si dice, con un neologismo brutto ma efficace, che uno schema campionario *implementa* il corrispondente disegno.

La motivazione principale che sottende l'uso di schemi campionari sta proprio nel fatto che consentono di selezionare un campione in base ad un dato disegno, e, soprattutto, sono facilmente realizzabili mediante programmi informatici. Per il momento non approfondiamo oltre il discorso; l'argomento verrà ripreso, ad un livello più avanzato, nei Capitoli 12 (dedicato ad aspetti generali relativi a disegni a probabilità variabili) e 15 (dedicato a specifici disegni campionari a probabilità variabili, di particolare utilità sul piano applicativo). Aspetti generali sugli schemi campionari, peraltro utili più da un punto di vista teorico che applicativo, sono nel volume Cassel *e altri* (1977) (pp. 15–16). Molto più moderno, ed anche molto più utile, è il volume di Tillé (2006), interamente dedicato alla costruzione di schemi di campionamento.

2.3 Dati campionari etichettati

Per ciascuna delle unità della popolazione selezionate nel campione $\mathbf{s}$, si osservano le corrispondenti modalità del carattere $\mathcal{Y}$ di interesse. Alla fine del processo di osservazione, si ottiene il *campione di modalità etichettate*, formato dalle coppie

$$(\textit{unità campionarie},\ \textit{modalità}).$$

Formalmente, il campione di modalità etichettate $\boldsymbol{y}(\mathbf{s})$ è costituito dalle coppie $(i,\ y_i)$, per tutte le unità i del campione $\mathbf{s}$. In simboli, si ha

$$\boldsymbol{y}(\mathbf{s}) = \{(i,\ y_i);\ i \in \mathbf{s}\}$$

se il campione $\mathbf{s}$ è un insieme, ed analoga espressione se è una sequenza (cioè se sono presenti ripetizioni e/o ordine). In ogni caso, $\boldsymbol{y}(\mathbf{s})$ contiene tutti i *dati statistici* ottenuti mediante la rilevazione campionaria.

Esempio 2.7. Si consideri il disegno campionario dell'Esempio 2.1, e si supponga che le modalità delle unità della popolazione siano le seguenti:

$$y_1 = 25,\ y_2 = 32,\ y_3 = 25,\ y_4 = 51,\ y_5 = 28,\ y_6 = 28,\ y_7 = 34.$$

Se si seleziona il campione $\mathbf{s}_1$, il corrispondente campione di modalità etichettate è

$$\boldsymbol{y}(\mathbf{s}_1) = ((1,\ 25),\ (2,\ 32),\ (3,\ 25)).$$

Se invece si seleziona $\mathbf{s}_6$, si ha

$$\boldsymbol{y}(\mathbf{s}_6) = ((1,\ 25),\ (2,\ 32),\ (2,\ 32),\ (3,\ 25)).$$

Si noti come le due modalità etichettate (1, 25) e (3, 25) siano da considerarsi differenti in quanto, pur essendo la stessa la modalità del carattere $\mathcal{Y}$ (25 in ambedue i casi) è diversa l'unità a cui questa si riferisce. □

È importante sottolineare che $\boldsymbol{y}(\mathbf{s})$ contiene le modalità campionarie *etichettate*: per ogni *modalità* y_i è presente anche l'*unità* i a cui essa si riferisce. In questo modo, nei dati campionari $\boldsymbol{y}(\mathbf{s})$ è sempre presente il collegamento tra unità campionate e modalità corrispondenti. Le modalità non sono semplici numeri (o attributi), ma recano con sé l'informazione relativa alle unità a cui si riferiscono.

Parallelamente alla riduzione del campione di unità, si può costruire il corrispondente *campione di modalità etichettate ridotto*, per il quale si userà la notazione $\boldsymbol{y}(r(\mathbf{s}))$:

$$\begin{aligned} \textit{campione di unità}:\ \mathbf{s} &\rightarrow \textit{riduzione}:\ r(\mathbf{s}) \\ \textit{campione di modalità etichettate}:\ \boldsymbol{y}(\mathbf{s}) &\rightarrow \textit{riduzione}:\ \boldsymbol{y}(r(\mathbf{s})). \end{aligned}$$

In altre parole, il campione di modalità etichettate ridotto è l'insieme delle modalità etichettate delle unità campionarie distinte. In simboli:

$$\boldsymbol{y}(r(\mathbf{s})) = \{(i,\, y_i);\ i \in r(\mathbf{s})\}.$$

Esempio 2.8. Si consideri ancora l'Esempio 2.7. Come già visto, è $\boldsymbol{y}(\mathbf{s}_6) = ((1,\, 25),\ (2,\, 32),\ (2,\, 32),\ (3,\, 25))$. Essendo poi la riduzione di $\mathbf{s}_6$ eguale a $r(\mathbf{s}_6) = \{1,\, 2,\, 3\}$, il corrispondente campione di modalità etichettate ridotto risulta: $\boldsymbol{y}(r(\mathbf{s}_6)) = \{(1,\, 25),\ (2,\, 32),\ (3,\, 25)\}$. Allo stesso modo, si vede facilmente che $\boldsymbol{y}(r(\mathbf{s}_1)) = \{(1,\, 25),\ (2,\, 32),\ (3,\, 25)\}$, in quanto i due campioni $\mathbf{s}_1$, $\mathbf{s}_6$ possiedono la stessa riduzione. □

Una questione importante riguarda l'eventuale differenza di contenuto informativo tra $\boldsymbol{y}(\mathbf{s})$ e $\boldsymbol{y}(r(\mathbf{s}))$. Il campione ridotto $\boldsymbol{y}(r(\mathbf{s}))$ differisce da quello "originale" $\boldsymbol{y}(\mathbf{s})$ solo perché in quest'ultimo vi sono unità osservate più volte (ripetizioni) e/o perché le unità stesse sono ordinate. Intuitivamente, l'osservare più volte la modalità di una stessa unità non porta nessuna informazione aggiuntiva rispetto all'osservarla una sola volta. Alla fine, ciò che *realmente* si osserva è soltanto l'unità con la corrispondente modalità. Pertanto, *le ripetizioni non portano nessuna informazione aggiuntiva.* Osservare una stessa unità una, due o più volte è perfettamente identico dal punto di vista dell'informazione che si ottiene. Una considerazione simile vale per l'ordine con cui le unità campionarie sono osservate. Osservare le stesse unità, ma con un ordine diverso, è equivalente dal punto di vista dell'informazione che si ottiene.

Il succo di quanto finora detto è che *ripetizioni e ordine sono irrilevanti* per quanto riguarda l'informazione fornita dai dati campionari. Ciò che è realmente importante sono le unità campionarie distinte e non ordinate, e le corrispondenti modalità. Formalmente, questo significa che il campione di modalità etichettate "originario" $\boldsymbol{y}(\mathbf{s})$ e la sua riduzione $\boldsymbol{y}(r(\mathbf{s}))$ *hanno lo stesso contenuto informativo.* L'utilizzare $\boldsymbol{y}(r(\mathbf{s}))$ come dati statistici in luogo di $\boldsymbol{y}(\mathbf{s})$ non porta a *nessuna perdita di informazione.* Poiché ripetizioni e ordine sono irrilevanti, d'ora in avanti si considereranno (quasi) esclusivamente disegni campionari senza ripetizioni e non ordinati.

Gli argomenti usati in questa sezione poggiano essenzialmente sull'intuizione. Su un piano più formale, il risultato fondamentale che giustifica quanto detto finora è che $\boldsymbol{y}(r(\mathbf{s}))$ è una statistica sufficiente minimale. Questo risultato verrà esposto più avanti, nella parte dedicata ai principî di inferenza da popolazioni finite (Capitolo 13).

2.4 Inferenza da popolazioni finite e inferenza da modello: due approcci a confronto

La teoria dell'inferenza da popolazioni finite presenta profonde differenze rispetto alla teoria dell'inferenza "da modello". Nella teoria dell'inferenza da modello il carattere $\mathcal{Y}$ è rappresentato da una variabile aleatoria a cui risulta

associata una distribuzione di probabilità $f(y;\theta)$ di forma nota dipendente da un parametro incognito θ. In tale impostazione le osservazioni campionarie $(y_1,\ldots,y_i,\ldots,y_n)$ rappresentano una realizzazione di una variabile casuale $(\mathcal{Y}_1,\ldots,\mathcal{Y}_i,\ldots,\mathcal{Y}_n)$ costituita da n variabili aleatorie indipendenti ed identicamente distribuite (*i.i.d.*) e con la stessa distribuzione di $\mathcal{Y}$. Assumendo l'ipotetica ripetibilità del processo di generazione dei dati sotto condizioni identiche, il campionamento avviene direttamente dal processo generatore dei dati stesso. Il riferimento è chiaramente ad una popolazione infinita. L'obiettivo è stimare il parametro incognito θ attraverso la definizione di una opportuna funzione delle osservazioni campionarie (denominata stimatore) $t(\mathcal{Y}_1,\ldots,\mathcal{Y}_i,\ldots,\mathcal{Y}_n)$.

Nella teoria dell'inferenza da popolazioni finite su cui si fonda il presente volume si assume che la popolazione sia composta da un numero finito di unità statistiche sulle quali sia possibile osservare il carattere di interesse $\mathcal{Y}$. I valori che $\mathcal{Y}$ assume sulle unità della popolazione sono quantità costanti e i parametri oggetto di inferenza sono valori sintetici che descrivono aspetti significativi del modo di manifestarsi del carattere nella popolazione (ad esempio, la media di $\mathcal{Y}$, la varianza di $\mathcal{Y}$, etc.). L'unica fonte di aleatorietà nella teoria dell'inferenza da popolazioni finite risiede nella probabilità che le unità della popolazione hanno di entrare a far parte del campione ossia nella probabilità con cui i diversi campioni della popolazione possono essere selezionati. Tali probabilità variano a seconda del disegno di campionamento adottato.

Anche in questo contesto l'obiettivo è stimare il parametro incognito attraverso uno stimatore, ma le sue proprietà si ricavano ipotizzando di poter selezionare dalla popolazione finita tutti i possibili campioni secondo il disegno di campionamento prefissato.

2.5 Stimatori e loro proprietà

Nei capitoli successivi ci si concentrerà principalmente sui problemi di stima della media della popolazione. Per questa ragione nella presente sezione si introduce brevemente la nozione di stimatore di un parametro di interesse, e se ne definiscono alcune proprietà.

Una *statistica campionaria* $T = t(\boldsymbol{y}(\mathbf{s}))$ è una funzione del campione di modalità etichettate (ossia dei dati statistici osservati su base campionaria). Il dominio di $t(\cdot)$ è, in generale, l'insieme $\{\boldsymbol{y}(\mathbf{s});\ \mathbf{s}\in\mathcal{S}\}$ di tutti i campioni di modalità osservabili. Poiché il campione di unità è essenzialmente una variabile aleatoria, che assume come valori i singoli $\mathbf{s}\in\mathcal{S}$ con probabilità $p(\mathbf{s})$, anche $T = t(\boldsymbol{y}(\mathbf{s}))$ è una variabile aleatoria. Esempi molto semplici di statistiche campionarie sono:

- $T = \sum_{i\in\mathbf{s}} y_i$ ammontare campionario (somma delle modalità delle unità campionarie);
- $T = \boldsymbol{y}(r(\mathbf{s}))$ campione ridotto di modalità etichettate;

- $T = \nu(\mathbf{s})$ ampiezza campionaria effettiva;
- $T = \max_{i \in \mathbf{s}} y_i$ massimo campionario.

In generale, il problema di *stima puntuale* che ci troviamo ad affrontare può essere descritto in modo molto semplice. Supponiamo che il parametro di interesse sia $\theta = \theta(\boldsymbol{Y}_N)$. Indichiamo poi con Θ l'insieme dei possibili valori che può assumere il parametro statistico $\theta(\boldsymbol{Y}_N)$, al variare di $\boldsymbol{Y}_N$ in Ω_N. Sulla base dei dati campionari, ossia sulla base del campione di modalità etichettate $\boldsymbol{y}(\mathbf{s})$, bisogna produrre una qualche ragionevole "approssimazione numerica" di θ. Tale obiettivo è raggiunto mediante l'uso di uno *stimatore*, ovvero di un'opportuna funzione dei dati campionari. Precisamente, uno stimatore $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y}(\mathbf{s}))$ di θ è una funzione dei dati campionari che ad ogni $\boldsymbol{y}(\mathbf{s})$ associa un possible valore dell'incognito parametro θ. In termini equivalenti, uno stimatore T di θ è una statistica campionaria a valori in Θ. Ovviamente, essendo il campione $\mathbf{s}$ aleatorio, lo stimatore $\widehat{\theta}$ è una variabile aleatoria.

Il termine stimatore non dovrebbe essere confuso con il termine stima, anche se spesso nella pratica i due termini sono utilizzati come sinonimi. Lo stimatore rappresenta una variabile aleatoria funzione dei dati campionari, il cui valore varia al variare del campione nello spazio campionario; la stima è il valore che lo stimatore assume in corrispondenza del campione osservato.

Se θ viene stimato con $\widehat{\theta}$, si commette in generale un *errore di stima*, pari a $\widehat{\theta} - \theta$. In assenza di altre fonti di errore e distorsioni, tale errore viene denominato errore campionario poiché deriva dalla parzialità dell'osservazione, cioè dalla circostanza che stiamo osservando solo una parte della popolazione. In corrispondenza dello specifico campione $\mathbf{s}$, si avrà un errore di stima pari a

$$\widehat{\theta}(\boldsymbol{y}(\mathbf{s})) - \theta(\boldsymbol{Y}_N).$$

Uno stimatore, ovviamente, è tanto migliore quanto più piccolo è l'errore di stima corrispondente. Poiché $\widehat{\theta}$ è una variabile aleatoria, l'errore di stima stesso è una variabile aleatoria. Per misurare la qualità di uno stimatore è quindi necessario adottare una qualche *misura di sintesi* dell'errore di stima, che:

- sia nulla se e solo se l'errore di stima è identicamente nullo per ogni campione $\mathbf{s}$ di $\mathcal{S}$;
- sia tanto più grande quanto più alta è la probabilità di campioni che portano a "grandi" errori di stima.

La misura di sintesi più usata, ed alla quale faremo sempre riferimento, è l'*errore quadratico medio* (*Mean Squared Error* nella terminologia anglosassone) di $\widehat{\theta}$:

$$MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = \sum_{\mathbf{s} \in \mathcal{S}} (\widehat{\theta}(\boldsymbol{y}(\mathbf{s})) - \theta(\boldsymbol{Y}_N))^2 \, p(\mathbf{s}). \tag{2.1}$$

Come messo in evidenza dalla (2.1), l'errore quadratico medio di uno stimatore dipende da tre elementi:

- la forma funzionale dello stimatore;
- il disegno campionario usato;
- le modalità y_i delle unità della popolazione, ovvero il parametro $\boldsymbol{Y}_N$ della popolazione.

Di questi tre elementi, i primi due sono scelti dallo statistico che progetta la rilevazione campionaria. Il terzo, invece, è fuori dal controllo dello statistico. In un certo senso, $\boldsymbol{Y}_N$ è deciso dalla natura. Il compito dello statistico è quindi quello di scegliere una coppia (disegno campionario, stimatore) che dia luogo ad un errore quadratico medio "piccolo", per quanto possibile. Si osservi che essendo il parametro della popolazione $\boldsymbol{Y}_N$ incognito, l'errore quadratico medio $MSE(\widehat{\theta})$ non può essere calcolato in pratica, numericamente.

In generale, non è difficile vedere dalla (2.1) che se si vuole stimare un parametro $\theta = \theta(\boldsymbol{Y}_N)$, e se $\widehat{\theta}$ è un suo stimatore, si ha

$$MSE(\widehat{\theta}) = 0 \; \textit{per ogni} \; \boldsymbol{Y}_N \in \Omega_N$$

soltanto se

$$\widehat{\theta}(\boldsymbol{y}(\mathbf{s})) = \theta(\boldsymbol{Y}_N) \; \textit{per ogni} \; \mathbf{s} \in \mathcal{S}, \; \boldsymbol{Y}_N \in \Omega_N \, .$$

Come conseguenza, si ha che (Esercizio 2.6) *non esiste uno stimatore il cui errore quadratico medio sia più piccolo di quello di ogni altro stimatore dello stesso parametro.*

In vista di questo risultato, e dell'ampiezza della classe di tutti gli stimatori di un parametro θ, un approccio molto naturale per la ricerca di un "buono" stimatore è quello di limitarsi ai soli stimatori che soddisfino qualche condizione aggiuntiva. La più semplice è che lo stimatore $\widehat{\theta}$ sia *corretto*, ossia che il suo valore atteso (sullo spazio dei campioni) sia esattamente uguale al parametro di interesse θ che si vuole stimare. In simboli:

$$E[\widehat{\theta}] = \theta(\boldsymbol{Y}_N)$$

qualunque sia il parametro della popolazione $\boldsymbol{Y}_N$, o, equivalentemente,

$$\sum_{\mathbf{s} \in \mathcal{S}} \widehat{\theta}(\boldsymbol{y}(\mathbf{s})) \; p(\mathbf{s}) = \theta(\boldsymbol{Y}_N)$$

qualunque sia $\boldsymbol{Y}_N$ in Ω_N.

Esattamente come nel caso dell'errore quadratico medio, il valore atteso di uno stimatore dipende da tre elementi: *a.* la forma funzionale dello stimatore; *b.* il disegno campionario usato; *c.* il parametro $\boldsymbol{Y}_N$ della popolazione. Da questa semplice osservazione discende che uno stesso stimatore può essere distorto se usato con un dato disegno, e corretto se usato con un differente disegno. La correttezza, quindi, non è una proprietà "intrinseca" di uno stimatore, ma è piuttosto legata alla coppia (*disegno campionario*, *stimatore*) che si utilizza. Tale coppia è in genere denominata *strategia di campionamento.*

Quanto detto significa che il giudizio su uno stimatore deve essere emesso valutando l'intera strategia campionaria (disegno campionario, stimatore), ossia la procedura che ha condotto alla formazione della stima ottenuta.

Esempio 2.9. Sia $I_4 = \{1, 2, 3, 4\}$ una popolazione di $N = 4$ unità, e sia $\boldsymbol{Y}_4$ il corrispondente vettore di modalità. Consideriamo poi il disegno campionario in cui lo spazio dei campioni è formato da:

$$\begin{aligned} &\mathbf{s}_1 = \{1, 2\},\ \mathbf{s}_2 = \{1, 3\},\ \mathbf{s}_3 = \{1, 4\},\ \mathbf{s}_4 = \{2, 3\}, \\ &\mathbf{s}_5 = \{2, 4\},\ \mathbf{s}_6 = \{3, 4\} \end{aligned} \tag{2.2}$$

con le seguenti probabilità:

$$p(\mathbf{s}_1) = p(\mathbf{s}_2) = \cdots = p(\mathbf{s}_6) = 1/6.$$

Come stimatore della media della popolazione, μ_y, consideriamo poi la media campionaria:

$$\overline{y}_{\mathbf{s}} = \frac{1}{2} \sum_{i \in \mathbf{s}} y_i$$

ossia la media delle modalità delle unità campionarie. Impiegato in coppia con il disegno sopra definito, lo stimatore $\overline{y}_{\mathbf{s}}$ è corretto. Infatti:

$$\begin{aligned} E\left[\overline{y}_{\mathbf{s}}\right] &= \sum_{\mathbf{s} \in \mathcal{S}} \overline{y}_{\mathbf{s}}\, p(\mathbf{s}) \\ &= \overline{y}_{\mathbf{s}_1}\, p(\mathbf{s}_1) + \cdots + \overline{y}_{\mathbf{s}_6}\, p(\mathbf{s}_6) \\ &= \frac{1}{6} \left\{ \frac{y_1 + y_2}{2} + \frac{y_1 + y_3}{2} + \frac{y_1 + y_4}{2} + \frac{y_2 + y_3}{2} + \frac{y_2 + y_4}{2} + \frac{y_3 + y_4}{2} \right\} \\ &= \frac{y_1 + y_2 + y_3 + y_4}{4} \\ &= \mu_y. \end{aligned}$$

Fermo restando lo stimatore $\overline{y}_{\mathbf{s}}$, consideriamo adesso un secondo disegno campionario, in cui lo spazio dei campioni è ancora (2.2), ma in cui i campioni hanno le seguenti probabilità:

$$p(\mathbf{s}_1) = p(\mathbf{s}_2) = p(\mathbf{s}_3) = 1/10,\ p(\mathbf{s}_4) = p(\mathbf{s}_5) = 2/10,\ p(\mathbf{s}_6) = 3/10.$$

In questo caso $\overline{y}_{\mathbf{s}}$ è distorto, in quanto:

$$\begin{aligned} E\left[\overline{y}_{\mathbf{s}}\right] &= \overline{y}_{\mathbf{s}_1}\, p(\mathbf{s}_1) + \cdots + \overline{y}_{\mathbf{s}_6}\, p(\mathbf{s}_6) \\ &= \frac{1}{10} \left\{ \frac{y_1 + y_2}{2} + \frac{y_1 + y_3}{2} + \frac{y_1 + y_4}{2} \right\} + \frac{2}{10} \left\{ \frac{y_2 + y_3}{2} + \frac{y_2 + y_4}{2} \right\} \\ &\quad + \frac{3}{10} \frac{y_3 + y_4}{2} \\ &= \frac{3y_1 + 5y_2 + 6y_3 + 6y_4}{20} \\ &\neq \mu_y. \end{aligned}$$

□

La *distorsione* di uno stimatore $\widehat{\theta}$, indicata con il simbolo $B(\widehat{\theta})$, è pari alla differenza tra il valore atteso di $\widehat{\theta}$ e il parametro θ da stimare. In simboli:

$$B(\widehat{\theta}) = E[\widehat{\theta}] - \theta.$$

La *varianza* di uno stimatore $\widehat{\theta}$, indicata con il simbolo $V(\widehat{\theta})$, è invece pari a:

$$V(\widehat{\theta}) = E\left[(\widehat{\theta} - E[\widehat{\theta}])^2\right] = \sum_{\mathbf{s}\in\mathcal{S}} (\widehat{\theta}(\boldsymbol{y}(\mathbf{s})) - E[\widehat{\theta}])^2 p(\mathbf{s}).$$

È appena il caso di menzionare che, per ben noti risultati, si può scrivere $V(\widehat{\theta}) = E[\widehat{\theta}^2] - (E[\widehat{\theta}])^2$.

Un risultato fondamentale riguarda la decomposizione dell'errore quadratico medio di uno stimatore $\widehat{\theta}$ nella somma di due termini: la varianza di $\widehat{\theta}$ e il quadrato della sua distorsione.

Proposizione 2.1. *Se $\widehat{\theta}$ è uno stimatore del parametro θ, si ha*

$$MSE(\widehat{\theta}) = V(\widehat{\theta}) + B(\widehat{\theta})^2\,. \tag{2.3}$$

Dimostrazione. In primo luogo, osserviamo che:

$$\begin{aligned} MSE(\widehat{\theta}) &= E[(\widehat{\theta} - \theta)^2] \\ &= E\left[\{(\widehat{\theta} - E[\widehat{\theta}]) + (E[\widehat{\theta}] - \theta)\}^2\right] \\ &= E\left[\{(\widehat{\theta} - E[\widehat{\theta}]) + B(\widehat{\theta})\}^2\right] \\ &= E\left[\{(\widehat{\theta} - E[\widehat{\theta}])^2 + B(\widehat{\theta})^2 + 2B(\widehat{\theta})\,(\widehat{\theta} - E[\widehat{\theta}])\right] \\ &= E\left[(\widehat{\theta} - E[\widehat{\theta}])^2\right] + E\left[B(\widehat{\theta})^2\right] + 2B(\widehat{\theta})\,E\left[\widehat{\theta} - E[\widehat{\theta}]\right]. \end{aligned} \tag{2.4}$$

Essendo poi $E[B(\widehat{\theta})^2] = B(\widehat{\theta})^2$ e $E[\widehat{\theta} - E[\widehat{\theta}]] = E[\widehat{\theta}] - E[\widehat{\theta}] = 0$, dalla (2.4) segue la (2.3). □

Come conseguenza dell'uguaglianza (2.3), si ha che *se uno stimatore è corretto, il suo errore quadratico medio si riduce alla sua varianza.*

Il restringersi a considerare solo stimatori corretti porta spesso a notevoli semplificazioni, ma può anche portare ad un prezzo da pagare in termini di efficienza: potrebbero esistere stimatori distorti con errore quadratico medio più piccolo di quello dello stimatore corretto che si considera. Inoltre, anche se uno stimatore è corretto le stime campionarie corrispondenti a tutti i possibili campioni selezionabili secondo il prefissato piano di campionamento differiranno (in positivo o in negativo) dal parametro di interesse θ. Quindi le stime presenteranno una variabilità intorno a θ. Se tale variabilità è elevata è del pari elevata la probabilità che la stima ottenuta da un campione casuale

risulti anche molto diversa dal parametro di interesse della popolazione. Al contrario, se la variabilità è piccola la distribuzione campionaria è non solo centrata ma anche concentrata intorno a θ. Di conseguenza, è alta la probabilità di selezionare casualmente campioni a cui corrispondono stime prossime a θ.

Esempio 2.10. Consideriamo ancora l'Esempio 2.9. Essendo $\overline{y}_{\mathbf{s}}$ corretto, il suo errore quadratico medio è uguale alla sua varianza: $MSE(\overline{y}_{\mathbf{s}}) = V(\overline{y}_{\mathbf{s}})$. Quest'ultima è pari a:

$$\begin{aligned}
V(\overline{y}_{\mathbf{s}}) &= E[\overline{y}_{\mathbf{s}}^2 - E[\overline{y}_{\mathbf{s}}])^2] \\
&= E[\overline{y}_{\mathbf{s}}^2] - E[\overline{y}_{\mathbf{s}}]^2 \\
&= \sum_{\mathbf{s}\in\mathcal{S}} \overline{y}_{\mathbf{s}}^2 - \mu_y^2 \\
&= \frac{1}{6}\left\{\left(\frac{y_1+y_2}{2}\right)^2 + \left(\frac{y_1+y_3}{2}\right)^2 + \left(\frac{y_1+y_4}{2}\right)^2 + \left(\frac{y_2+y_3}{2}\right)^2\right. \\
&\quad \left. + \left(\frac{y_2+y_4}{2}\right)^2 + \left(\frac{y_3+y_4}{2}\right)^2\right\} - \left(\frac{y_1+y_2+y_3+y_4}{4}\right)^2 \\
&= \frac{1}{6}\left\{\frac{3}{4}(y_1^2+y_2^2+y_3^2+y_4^2)\right. \\
&\quad \left. +2\left(\frac{y_1y_2+y_1y_3+y_1y_4+y_2y_3+y_2y_4+y_3y_4}{4}\right)\right\} \\
&\quad -\frac{y_1^2+y_2^2+y_3^2+y_4^2}{16} - \frac{y_1y_2+y_1y_3+y_1y_4+y_2y_3+y_2y_4+y_3y_4}{8} \\
&= \frac{1}{3}\left\{\frac{y_1^2+y_2^2+y_3^2+y_4^2}{4} - \left(\frac{y_1+y_2+y_3+y_4}{4}\right)^2\right\} \\
&= \frac{\sigma_y^2}{3}
\end{aligned}$$

essendo

$$\sigma_y^2 = \frac{1}{4}\sum_{i=1}^{4}(y_i-\mu_y)^2 = \frac{y_1^2+y_2^2+y_3^2+y_4^2}{4} - \left(\frac{y_1+y_2+y_3+y_4}{4}\right)^2$$

la varianza della popolazione. □

Esempio 2.11. Si consideri l'Esempio 2.10. Lo stimatore $\overline{y}_{\mathbf{s}}$ è ora distorto, con distorsione pari a:

$$B(\overline{y}_{\mathbf{s}}) = \frac{3y_1+5y_2+6y_3+6y_4}{20} - \frac{y_1+y_2+y_3+y_4}{4} = \frac{-2y_1+y_3+y_4}{20}. \quad (2.5)$$

La varianza di $\overline{y}_{\mathbf{s}}$ è invece eguale a:

$$\begin{aligned} V(\overline{y}_{\mathbf{s}}) &= E[\overline{y}_{\mathbf{s}}^2] - E[\overline{y}_{\mathbf{s}}]^2 = \sum_{\mathbf{s}\in\mathcal{S}} \overline{y}_{\mathbf{s}}^2 - \mu_y^2 \\ &= \frac{1}{10}\left(\frac{y_1+y_2}{2}\right)^2 + \frac{1}{10}\left(\frac{y_1+y_3}{2}\right)^2 + \frac{1}{10}\left(\frac{y_1+y_4}{2}\right)^2 \\ &\quad + \frac{2}{10}\left(\frac{y_2+y_3}{2}\right)^2 + \frac{2}{10}\left(\frac{y_2+y_4}{2}\right)^2 + \frac{3}{10}\left(\frac{y_3+y_4}{2}\right)^2 \\ &\quad - \left(\frac{3y_1+5y_2+6y_3+6y_4}{20}\right)^2. \end{aligned} \tag{2.6}$$

La somma di (2.6) e del quadrato di (2.5) fornisce l'errore quadratico medio di $\overline{y}_{\mathbf{s}}$. □

L'errore quadratico medio $MSE(\widehat{\theta})$ di uno stimatore $\widehat{\theta}$ (o la varianza $V(\widehat{\theta})$ nel caso in cui $\widehat{\theta}$ sia corretto) misura l'imprecisione di $\widehat{\theta}$ in termini *assoluti*. Spesso è importante misurare tale imprecisione in termini *relativi*, calcolando il rapporto percentuale tra $\sqrt{MSE(\widehat{\theta})}$ e il valore (assoluto) del parametro che si vuole stimare:

$$\frac{\sqrt{MSE(\widehat{\theta})}}{|\theta|}\,100. \tag{2.7}$$

In particolare, quando lo stimatore $\widehat{\theta}$ è corretto si ha $\theta = E[\widehat{\theta}]$ e $MSE(\widehat{\theta}) = V(\widehat{\theta})$, per cui la (2.7) si riduce al *coefficiente di variazione* di $\widehat{\theta}$. In simboli:

$$CV(\widehat{\theta}) = \frac{\sqrt{V(\widehat{\theta})}}{|E[\widehat{\theta}]|}\,100$$

purché il valore atteso al denominatore non sia nullo.

La selezione di un campione di unità, l'osservazione delle corrispondenti modalità, e la costruzione di stimatori dei parametri di interesse *non* esauriscono il lavoro dello statistico. Infatti, ogni stima di un parametro va *sempre* accompagnata da una stima del suo grado di "bontà", di "affidabilità". Come detto dianzi, la principale misura di quanto "buono" o "cattivo" sia uno stimatore è costituita dal suo errore quadratico medio, il quale dipende dall'intero vettore $\boldsymbol{Y}_N$ delle modalità di tutte le unità della popolazione. Non è quindi possibile calcolare realmente il valore che esso assume. Lo sarebbe solo se $\boldsymbol{Y}_N$ fosse noto, il che non accade mai nella pratica applicativa. In effetti, se $\boldsymbol{Y}_N$ fosse noto non ci sarebbe nessun motivo di ricorrere ad una rilevazione campionaria.

Se non è possibile dire quale sia il valore dell'errore quadratico medio di uno stimatore, è però in genere possibile *stimarlo* sulla base dei dati campionari disponibili. Pertanto, ogni volta che si costruisce una stima di un parametro incognito, bisogna anche produrre una stima del suo errore quadratico medio.

Se poi lo stimatore che si utilizza è corretto, il suo errore quadratico medio coincide con la sua varianza, e quindi stimare il suo errore quadratico medio equivale a stimare la sua varianza.

Quanto detto nella presente sezione rende chiaro il ruolo centrale svolto, nell'ambito del campionamento da popolazioni finite, dalle regole di selezione dei campioni di tipo probabilistico, ossia dai disegni campionari. In effetti, solo quando la scelta del campione è attuata mediante un disegno campionario è possibile studiare in termini quantitativi, precisi, il comportamento di uno stimatore. Solo in questo caso, infatti, ha senso il calcolo del suo errore quadratico medio.

Quando la scelta del campione di unità non è attuata mediante un disegno campionario controllato dallo statistico, non si è in grado di assegnare un valore alle probabilità $p(\mathbf{s})$ dei diversi campioni, e spesso non si è neanche in grado di elencare, neppure in linea puramente concettuale, tutti i possibili campioni dell'insieme $\mathcal{S}$. In questi casi perde di significato il riferirsi all'errore quadratico medio di uno stimatore come misura di quanto "buono" o "cattivo" esso sia. Infatti, quando o non si è in grado di esplicitare lo spazio $\mathcal{S}$ dei campioni, o quando non si conoscono i valori assunti dalle probabilità $p(\mathbf{s})$ dei diversi campioni, l'errore quadratico medio di uno stimatore non è calcolabile né stimabile, neppure nel caso (che ovviamente non si verifica mai nella pratica applicativa) in cui siano note le modalità y_i delle unità della popolazione.

Queste considerazioni chiariscono a sufficienza il perché, nella presente trattazione, verrà data la massima enfasi alle regole di selezione di campioni di unità basate su disegni campionari. Esse sono le sole regole di selezione *controllate dallo statistico*, e quindi le sole regole di selezione per le quali ha senso studiare le *performance* di stimatori di parametri di interesse e stimarne l'errore quadratico medio.

Nella pratica applicativa vengono a volte (in effetti abbastanza spesso) usate regole di selezione dei campioni *non* controllate dallo statistico. Un esempio molto semplice sono i sondaggi di opinione effettuati in trasmissioni televisive, in cui si invitano i telespettatori a telefonare ad un dato numero (o a inviare un sms, o un messaggio di e-*mail*), e a rispondere ad una o più domande. In questi casi le unità del campione si *autoselezionano*. Esse non sono scelte mediante una procedura controllata dello statistico, e non si è in grado di dire né quali siano i possibili campioni osservabili (lo spazio dei campioni), né quali siano le loro probabilità. In tali casi, a rigore non si è in grado di dire *nulla* sull'errore quadratico medio di stimatori di parametri di interesse, e quindi le stime campionarie hanno un valore molto limitato, quasi nullo. *Solo* stime campionarie ottenute mediante regole di selezione delle unità controllate dallo statistico (ossia mediante veri disegni campionari) hanno valore sostanziale. Ciò accade, come già affermato, per una ragione molto semplice. Solo per esse è possibile:

- studiare le proprietà di stimatori di parametri di interesse;
- stimare i loro errori quadratici medi.

2.6 Intervalli di confidenza

Accanto a stime *puntuali* come quelle brevemente esposte nella sezione precedente, è spesso di interesse fornire delle *stime intervallari*, in cui non si costruisce un'unica stima di un parametro di interesse, ma un intervallo di "plausibili stime" del parametro stesso.

In generale, siano $T_1 = t_1(\boldsymbol{y}(\mathbf{s}))$, $T_2 = t_2(\boldsymbol{y}(\mathbf{s}))$ due statistiche campionarie, tali che $t_1(\boldsymbol{y}(\mathbf{s})) \leqslant t_2(\boldsymbol{y}(\mathbf{s}))$ qualunque sia il campione (di modalità) $\boldsymbol{y}(\mathbf{s})$. Ha senso in questo caso considerare l'intervallo $[T_1,\, T_2]$. I suoi estremi sono, come detto, funzioni dei dati campionari. Si tratta quindi di due variabili aleatorie, il che giustifica il riferirsi a $[T_1,\, T_2]$ come ad un *intervallo aleatorio*. Diremo che $[T_1,\, T_2]$ è un *intervallo di confidenza al livello* $1-\alpha$ se esso contiene il parametro di interesse $\theta = \theta(\boldsymbol{Y}_N)$ con probabilità $1-\alpha$, qualunque sia il valore di θ, cioè qualunque sia $\boldsymbol{Y}_N$ in Ω_N. Ora, la probabilità che l'intervallo aleatorio $[T_1,\, T_2]$ racchiuda $\theta(\boldsymbol{Y}_N)$ è pari alla somma delle probabilità di tutti i campioni $\mathbf{s}$ tali che $t_1(\boldsymbol{y}(\mathbf{s})) \leqslant \theta(\boldsymbol{Y}_N) \leqslant t_2(\boldsymbol{y}(\mathbf{s}))$. In simboli, detto $E = \{\mathbf{s} \in \mathcal{S} :\ t_1(\boldsymbol{y}(\mathbf{s})) \leqslant \theta(\boldsymbol{Y}_N) \leqslant t_2(\boldsymbol{y}(\mathbf{s}))\}$ l'insieme dei campioni (di unità) tali che $\theta(\boldsymbol{Y}_N)$ è racchiuso nell'intervallo di estremi $t_1(\boldsymbol{y}(\mathbf{s}))$, $t_2(\boldsymbol{y}(\mathbf{s}))$, si può formalmente scrivere

$$Pr(T_1 \leqslant \theta(\boldsymbol{Y}_N) \leqslant T_2) = \sum_{\mathbf{s}\in E} p(\mathbf{s}). \tag{2.8}$$

Dalla (2.8) si desume anche che $[T_1,\, T_2]$ è un intervallo di confidenza al livello $1-\alpha$ per $\theta = \theta(\boldsymbol{Y}_N)$ se

$$\sum_{\mathbf{s}\in E} p(\mathbf{s}) = 1-\alpha \ \ \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N.$$

Spesso, i due estremi T_1, T_2 sono costruiti a partire da uno stimatore $\widehat{\theta}$ di θ. Supponiamo che θ sia uno stimatore corretto di $\theta = \theta(\boldsymbol{Y}_N)$, e siano poi q_1, q_2 due quantità tali che

$$Pr\left(q_1 \leqslant \widehat{\theta} - \theta(\boldsymbol{Y}_N) \leqslant q_2\right) = 1-\alpha \ \ \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N. \tag{2.9}$$

La (2.9) si può riscrivere come

$$Pr\left(\widehat{\theta} - q_2 \leqslant \theta(\boldsymbol{Y}_N) \leqslant \widehat{\theta} - q_1\right) = 1-\alpha \ \ \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N$$

e quindi, per $T_1 = \widehat{\theta} - q_2$, $T_2 = \widehat{\theta} - q_1$, si vede subito che $[\widehat{\theta} - q_2,\ \widehat{\theta} - q_1]$ è un intervallo di confidenza per θ al livello $1-\alpha$.

In generale, le due quantità q_1, q_2 che compaiono nella (2.9) sono determinate dalla distribuzione di probabilità dello stimatore $\widehat{\theta}$, la quale dipende da $\boldsymbol{Y}_N$. Pertanto gli stessi q_1, q_2 devono dipendere da $\boldsymbol{Y}_N$, e non sono di conseguenza calcolabili. Una soluzione a questa *impasse* si basa sull'idea di

approssimare la distribuzione di probabilità dello stimatore $\widehat{\theta}$. L'approssimazione più semplice e utile è quella *normale*. Se $V(\widehat{\theta})$ indica la varianza di $\widehat{\theta}$, la distribuzione di probabilità di

$$\frac{\widehat{\theta} - \theta}{\sqrt{V(\widehat{\theta})}}$$

viene approssimata con una normale standard, $N(0, 1)$. Non discutiamo qui della validità e dei limiti di questa approssimazione. Ad ogni modo, indicando con $z_{\alpha/2}$ il valore tale che $Pr(N(0, 1) \geqslant z_{\alpha/2}) = \alpha/2$, si ha in via approssimata

$$Pr\left(-z_{\alpha/2} \leqslant \frac{\widehat{\theta} - \theta}{\sqrt{V(\widehat{\theta})}} \leqslant z_{\alpha/2}\right) = 1 - \alpha$$

ovvero

$$Pr\left(\widehat{\theta} - z_{\alpha/2}\sqrt{V(\widehat{\theta})} \leqslant \theta(\boldsymbol{Y}_N) \leqslant \widehat{\theta} + z_{\alpha/2}\sqrt{V(\widehat{\theta})}\right) = 1 - \alpha \text{ } \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N.$$

Quindi, l'intervallo

$$\left[\widehat{\theta} - z_{\alpha/2}\sqrt{V(\widehat{\theta})},\ \widehat{\theta} + z_{\alpha/2}\sqrt{V(\widehat{\theta})}\right] \tag{2.10}$$

è un intervallo di confidenza (approssimato) per θ al livello $1 - \alpha$.

La varianza $V(\widehat{\theta})$, come detto più volte nella Sezione 2.4, dipende dall'intero vettore $\boldsymbol{Y}_N$ delle modalità di tutte le unità della popolazione, e quindi, a meno di casi eccezionali, è incognita. Un'idea naturale è ovviamente quella di stimarla sulla base dei dati campionari. Se $\widehat{V}(\widehat{\theta})$ è un opportuno stimatore di $V(\widehat{\theta})$, l'idea di base consiste nell'approssimare la distribuzione di probabilità di

$$\frac{\widehat{\theta} - \theta}{\sqrt{\widehat{V}(\widehat{\theta})}} \tag{2.11}$$

con una normale standard, $N(0, 1)$. Il sottrarre allo stimatore $\widehat{\theta}$ il suo valore atteso θ, e il dividere il tutto per $\sqrt{\widehat{V}(\widehat{\theta})}$ è in genere detto *studentizzazione* di $\widehat{\theta}$. Sostituendo $V(\widehat{\theta})$ con $\widehat{V}(\widehat{\theta})$ nella (2.10), si ha che

$$\left[\widehat{\theta} - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\theta})},\ \widehat{\theta} + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\theta})}\right] \tag{2.12}$$

è ancora un intervallo di confidenza approssimato per θ al livello $1 - \alpha$. La frase "$\widehat{V}(\widehat{\theta})$ sia un opportuno stimatore di $V(\widehat{\theta})$" è imprecisa e vaga. Nei casi più importanti vedremo di volta in volta la forma che assume $\widehat{V}(\widehat{\theta})$.

Esercizi

2.1. Sia $I_6 = \{1, \ldots, 6\}$ una popolazione finita di $N = 6$ unità, e si consideri il seguente disegno campionario:

$$\begin{aligned}
&\mathbf{s}_1 = (1, 1),\ \mathbf{s}_2 = (1, 5, 6),\ \mathbf{s}_3 = (1, 1, 2, 1),\ \mathbf{s}_4 = (6, 5, 1),\\
&\mathbf{s}_5 = (2, 1, 3),\ \mathbf{s}_6 = (2, 3, 1, 2),\ \mathbf{s}_7 = (4, 3, 1),\ \mathbf{s}_8 = (2, 3),\\
&\mathbf{s}_9 = (3, 2),\ \mathbf{s}_{10} = (1, 4, 3),\ \mathbf{s}_{11} = (3, 4, 1),\ \mathbf{s}_{12} = (1, 1, 1);\\
&p(\mathbf{s}_1) = p(\mathbf{s}_2) = \cdots = p(\mathbf{s}_{12}) = \frac{1}{12}.
\end{aligned}$$

a. Calcolare l'ampiezza media e l'ampiezza media effettiva di questo disegno.
b. Costruire la sua riduzione.
c. Costruire un disegno campionario ordinato e senza ripetizioni la cui riduzione sia identica a quella del disegno dell'esercizio.

2.2. Provare che se $(\mathcal{S}, p)$ è un disegno campionario e $(\mathcal{S}^*, p^*)$ è la sua riduzione, l'ampiezza media effettiva di $(\mathcal{S}, p)$ è uguale all'ampiezza media di $(\mathcal{S}^*, p^*)$.

Suggerimento. Tenere conto che se $\mathbf{s}^* \in \mathcal{S}^*$, allora $p^*(\mathbf{s}^*) = \sum_{\mathbf{s}\in C(\mathbf{s}^*)} p(\mathbf{s})$, e che valgono le ralazioni $\sum_{\mathbf{s}^*\in\mathcal{S}^*} n(\mathbf{s}^*)p^*(\mathbf{s}^*) = \sum_{\mathbf{s}^*\in\mathcal{S}^*} n(\mathbf{s}^*)\left\{\sum_{\mathbf{s}\in C(\mathbf{s}^*)} p(\mathbf{s})\right\}$ e $n(\mathbf{s}^*) = \nu(\mathbf{s})$.

2.3. Costruire la riduzione del disegno campionario dell'Esempio 2.4.

2.4. Con riferimento all'Esempio 2.7 (vds. anche Esempio 2.1), determinare i campioni di modalità etichettate $\boldsymbol{y}(\mathbf{s}_1), \ldots, \boldsymbol{y}(\mathbf{s}_8)$.

2.5. Con riferimento all'Esempio 2.7, determinare i campioni di modalità etichettate ridotti $\boldsymbol{y}(r(\mathbf{s}_1)), \ldots, \boldsymbol{y}(r(\mathbf{s}_8))$.

2.6. Preso un arbitrario vettore $\boldsymbol{Z}_N \in \Omega_N$, costruire lo stimatore di $\theta = \theta(\boldsymbol{Y}_N)$:

$$\widehat{\theta}_Z = \theta(\boldsymbol{Z}_N)\ \forall\, \mathbf{s} \in \mathcal{S}.$$

a. Verificare che $MSE(\widehat{\theta}_Z) = 0$ se $\boldsymbol{Y}_N = \boldsymbol{Z}_N$.
b. Dedurre dal punto *a.* che se $\widehat{\theta}^*$ è uno stimatore di θ con errore quadratico medio più piccolo di quello di ogni altro stimatore di θ, deve essere $MSE(\widehat{\theta}^*) = 0$ qualunque sia $\boldsymbol{Y}_N \in \Omega_N$.

Suggerimento. Se $MSE(\widehat{\theta}^*) \leqslant MSE(\widehat{\theta}_Z)$ qualunque sia $Y_N \in \Omega_N$, per $Y_N = Z_N$ si deve avere $MSE(\widehat{\theta}^*) = 0$. Ripetendo il ragionamento per tutti i vettori Z_N in Ω_N, cioè per ogni possibile stimatore $\widehat{\theta}_Z$ che è possibile costruire, si ha il risultato.

c. Dedurre dal punto *b.* che non esiste uno stimatore di θ con errore quadratico medio più piccolo di quello di ogni altro stimatore del parametro stesso, qualunque sia $\boldsymbol{Y}_N \in \Omega_N$.

2.7. Data una popolazione finita di $N = 4$ unità, e posto $\boldsymbol{Y}_4 = (y_1, y_2, y_3, y_4)$, si supponga di voler stimare la media $\mu_y = (y_1 + y_2 + y_3 + y_4)/4$. Il disegno campionario $(\mathcal{S}, p)$ che si utilizza è tale che lo spazio dei campioni è formato da:

$$\mathbf{s}_1 = \{1, 2, 3\},\ \mathbf{s}_2 = \{1, 2, 4\},\ \mathbf{s}_3 = \{2, 3, 4\},\ \mathbf{s}_4 = \{1, 3, 4\}$$

con probabilità

$$p(\mathbf{s}_1) = 0.25,\ p(\mathbf{s}_2) = 0.25,\ p(\mathbf{s}_3) = 0.2,\ p(\mathbf{s}_4) = 0.3.$$

Si considerino poi i due stimatori t_1, t_2 di μ_y definiti nel modo seguente:

$$\begin{aligned}
t_1(\boldsymbol{y}(\mathbf{s}_1)) &= \frac{y_1 + y_2 + y_3}{3},\ t_1(\boldsymbol{y}(\mathbf{s}_2)) = \frac{y_1 + y_2 + y_4}{3},\\
t_1(\boldsymbol{y}(\mathbf{s}_3)) &= \frac{y_2 + y_3 + y_4}{3},\ t_1(\boldsymbol{y}(\mathbf{s}_4)) = \frac{y_1 + y_3 + y_4}{3};\\
t_2(\boldsymbol{y}(\mathbf{s}_1)) &= \frac{y_1 + 2y_2 + 3y_3}{6},\ t_2(\boldsymbol{y}(\mathbf{s}_2)) = \frac{y_1 + 2y_2 + 4y_4}{7},\\
t_2(\boldsymbol{y}(\mathbf{s}_3)) &= \frac{2y_2 + 3y_3 + 4y_4}{9},\ t_2(\boldsymbol{y}(\mathbf{s}_4)) = \frac{y_1 + 3y_3 + 4y_4}{8}.
\end{aligned}$$

Calcolare $E[t_1]$, $V(t_1)$, $MSE(t_1)$, $E[t_2]$, $V(t_2)$, $MSE(t_2)$.

2.8. Data una popolazione finita di $N = 20$ unità, si consideri un disegno campionario in cui lo spazio dei campioni è formato dai campioni:

$$\begin{aligned}
&\mathbf{s}_1 = (1, 11),\ \mathbf{s}_2 = (2, 12),\ \mathbf{s}_3 = (3, 13),\ \mathbf{s}_4 = (4, 14),\ \mathbf{s}_5 = (5, 15),\\
&\mathbf{s}_6 = (6, 16),\ \mathbf{s}_7 = (7, 17),\ \mathbf{s}_8 = (8, 18),\ \mathbf{s}_9 = (9, 19),\ \mathbf{s}_{10} = (10, 20)
\end{aligned}$$

con probabilità

$$\begin{aligned}
&p(\mathbf{s}_1) = 0.1,\ p(\mathbf{s}_2) = 0.1,\ p(\mathbf{s}_3) = 0.025,\ p(\mathbf{s}_4) = 0.025,\ \mathbf{s}_5 = 0.25,\\
&p(\mathbf{s}_6) = 0.25,\ p(\mathbf{s}_7) = 0.025,\ p(\mathbf{s}_8) = 0.025,\ p(\mathbf{s}_9) = 0.1,\ p(\mathbf{s}_{10}) = 0.1.
\end{aligned}$$

Costruire, in base a quanto svolto nella Sezione 2.2, uno schema che implementa questo disegno.

3

Disegno campionario semplice

3.1 Il disegno semplice senza ripetizione

3.1.1 Definizione del disegno semplice senza ripetizione

Il disegno campionario *semplice senza ripetizione* (disegno ssr, d'ora in avanti) è probabilmente il più importante tra tutti i disegni campionari, sia per l'utilizzo diretto che se ne fa, sia perché entra come costituente essenziale di disegni campionari più complessi, (disegno stratificato, disegno a due stadi, etc.).

Sia $I_N = \{1, \ldots, N\}$ una popolazione finita di N unità. Il disegno ssr di numerosità n $(1 \leqslant n \leqslant N)$ è definito come segue:

- lo spazio dei campioni è l'insieme di tutte le combinazioni senza ripetizioni di classe n delle N unità della popolazione (ogni campione è uno dei possibili sottoinsiemi di n delle N unità della popolazione);
- tutti i campioni hanno la stessa probabilità di essere selezionati.

Formalmente, detto $\mathcal{C}_{N,n}$ l'insieme di tutte le combinazioni senza ripetizione di n delle N unità della popolazione, e tenendo conto che vi sono in totale $\binom{N}{n}$ di tali combinazioni, si ha:

$$\mathcal{S} = \mathcal{C}_{N,n}; \quad p(\mathbf{s}) = \frac{1}{\binom{N}{n}} \textit{ per ogni } \mathbf{s} \in \mathcal{C}_{N,n}.$$

Usando le definizioni date nella sezione precedente, si vede subito che il disegno ssr è *non ordinato*, *senza ripetizioni*, e ad *ampiezza effettiva costante* n.

Esempio 3.1. Per una popolazione di $N = 5$ unità, $I_5 = \{1, \ldots, 5\}$, consideriamo un disegno ssr di numerosità $n = 3$. Lo spazio dei campioni è formato dai $\binom{5}{3} = 10$ seguenti campioni:

$\mathbf{s}_1 = \{1, 2, 3\}$, $\mathbf{s}_2 = \{1, 2, 4\}$, $\mathbf{s}_3 = \{1, 2, 5\}$, $\mathbf{s}_4 = \{1, 3, 4\}$, $\mathbf{s}_5 = \{1, 3, 5\}$,
$\mathbf{s}_6 = \{1, 4, 5\}$, $\mathbf{s}_7 = \{2, 3, 4\}$, $\mathbf{s}_8 = \{2, 3, 5\}$, $\mathbf{s}_9 = \{2, 4, 5\}$, $\mathbf{s}_{10} = \{3, 4, 5\}$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_3,

ciascuno dei quali ha probabilità

$$p(\mathbf{s}_1) = \cdots = p(\mathbf{s}_{10}) = \frac{1}{\binom{5}{3}} = \frac{1}{10}. \qquad \square$$

3.1.2 Simmetria totale del disegno semplice senza ripetizione

La proprietà probabilmente più importante del disegno ssr è la sua *simmetria totale*. Per introdurre questa nozione, partiamo da un facile esempio.

Esempio 3.2. Due statistici devono estrarre un campione ssr di $n = 2$ unità da una popolazione di $N = 3$ individui, Antonio (A), Bruno (B), Carlo (C). Il primo statistico assegna a Antonio l'etichetta 1, a Bruno l'etichetta 2, a Carlo l'etichetta 3, e costruisce il corrispondente disegno ssr. Il secondo statistico, invece, assegna a Carlo l'etichetta 1, a Antonio l'etichetta 2, a Bruno l'etichetta 3, e costruisce anch'egli il corrispondente disegno ssr. I due disegni campionari sono mostrati nella Tabella 3.1.

Tabella 3.1 Disegni campionari costruiti dai due statistici

Disegno 1			*Disegno 2*		
Campioni		*Probabilità*	*Campioni*		*Probabilità*
Etichette	*Unità*		*Etichette*	*Unità*	
{1, 2}	{*A*, *B*}	1/3	{1, 2}	{*A*, *C*}	1/3
{1, 3}	{*A*, *C*}	1/3	{1, 3}	{*B*, *C*}	1/3
{2, 3}	{*B*, *C*}	1/3	{2, 3}	{*A*, *B*}	1/3

I due disegni campionari sono identici, per cui il diverso modo in cui i due statistici hanno assegnato le etichette alle unità della popolazione non ha avuto alcuna influenza sul disegno campionario. $\square$

Quanto descritto nell'Esempio 3.1 vale del tutto in generale: il disegno ssr *non dipende dal modo in cui le etichette sono assegnate alle unità della popolazione*. In qualunque modo si effettui l'assegnazione delle etichette alle unità, il disegno ssr rimane invariato. Poiché ogni possibile modo di *assegnare* le etichette $\{1, \ldots, N\}$ alle unità della popolazione equivale a *permutare* le etichette stesse, si può affermare, in termini equivalenti, che il disegno ssr resta identico comunque si permutino le etichette delle unità della popolazione. Questa caratteristica, che denomineremo *simmetria totale*, è la proprietà più importante del disegno ssr. Essa ci dice, in buona sostanza, che l'assegnare ad un'unità l'una o l'altra delle etichette $\{1, \ldots, N\}$, non altera il modo in cui l'unità stessa è trattata. L'ovvia conclusione è che il disegno ssr *tratta allo stesso modo tutte le unità della popolazione*.

3.1.3 Implementazione del disegno semplice senza ripetizione

La selezione di un campione ssr è facile da realizzare, e quasi tutti i più comuni *package* statistici (R, S-PLUS, SPSS, etc.) contengono procedure per l'implementazione del disegno ssr. Qui di seguito sono forniti due semplici algoritmi per implementare il disegno ssr, molto facili da usare.

Algoritmo 1.

- *Passo* 1. Generare N numeri aleatori $U_1, \ldots, U_N$ (uno per ogni unità della popolazione) con distribuzione uniforme in $[0, 1]$. Andare al Passo 2.
- *Passo* 2. Ordinare le N unità della popolazione da quella con il valore U più piccolo a quella con il valore U più grande. Andare al Passo 3.
- *Passo* 3. Prendere le n unità corrispondenti agli n valori U più piccoli. Esse formano un campione ssr di n unità della popolazione.

Algoritmo 2.

Definire N numeri $B_1, \ldots, B_N$ e due numeri (interi) t, i. Se x è un numero reale, porre $\lceil x \rceil =$ *più piccolo numero intero* $\geqslant x$.

- *Passo* 0. Inizializzazione. Porre $B_1 = 0, \ldots, B_N = 0$, $i = 0$, $t = 0$. Andare al Passo 1.
- *Passo* 1. Se $t = n$ andare al Passo 3. Altrimenti, generare un numero U aleatorio con distribuzione uniforme in $[0, 1]$, porre $i = \lceil N\,U \rceil$ e andare al Passo 2.
- *Passo* 2. Se $B_i = 0$ porre $B_i = 1$, incrementare t di 1 e tornare al Passo 1. Se invece $B_i = 1$, tornare al Passo 1.
- *Passo* 3. Arresto. Le unità del campione sono quelle con etichette corrispondenti agli indici i tali che $B_i = 1$.

3.2 Stima della media della popolazione: la media campionaria

Il problema principale che affronteremo in questo capitolo è la stima della media della popolazione:

$$\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Sia

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu_y)^2 = \frac{1}{N} \sum_{i=1}^{N} y_i^2 - \mu_y^2$$

la varianza della popolazione, e sia

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2 = \frac{N}{N-1}\sigma_y^2$$

la varianza della popolazione "corretta" con il denominatore $N-1$. Si osservi che per N abbastanza grande (dell'ordine di poche centinaia, o poche migliaia, di unità) è $(N-1)/N \approx 1$, per cui è $S_y^2 \approx \sigma_y^2$. A meno di casi eccezionali, nella pratica applicativa il sostituire S_y^2 con σ_y^2 è virtualmente privo di effetti.

Indichiamo al solito con $\boldsymbol{y}(\mathbf{s}) = \{y_i;\ i \in \mathbf{s}\}$ il campione di modalità etichettate (dati campionari). Il modo più naturale per definire uno stimatore di un parametro incognito $\theta = \theta(\boldsymbol{Y}_N)$ è quello di applicare alle n osservazioni campionarie y_i la stessa funzione $\theta(\cdot)$ che definisce il parametro di interesse nella popolazione. Resta in tal modo definito lo "stimatore conforme" di θ. Di conseguenza, lo stimatore conforme di μ_y è la *media campionaria*:

$$\overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} y_i.$$

In questa sezione sono studiate con un certo dettaglio le proprietà della media campionaria, ed in particolare la sua media e la sua varianza. Per facilitare tale studio, introduciamo la funzione indicatrice di presenza/assenza dell'unità i $(\in I_N)$ nel campione $\mathbf{s}$ $(\in \mathcal{C}_{N,n})$:

$$\delta(i;\, \mathbf{s}) = \begin{cases} 1 \text{ } se \text{ } i \in \mathbf{s} \\ 0 \text{ } se \text{ } i \notin \mathbf{s} \end{cases}.$$

Con questa convenzione, si può scrivere

$$\overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i=1}^{N} y_i\, \delta(i;\, \mathbf{s}). \tag{3.1}$$

Nella successiva proposizione viene mostrato che il valore atteso di $\overline{y}_{\mathbf{s}}$ è uguale alla media della popolazione, ovvero che $\overline{y}_{\mathbf{s}}$ è uno stimatore corretto di μ_y.

Proposizione 3.1. *Se il disegno campionario è ssr, la media campionaria è uno stimatore corretto della media della popolazione:*

$$E\left[\overline{y}_{\mathbf{s}}\right] = \mu_y. \tag{3.2}$$

Dimostrazione. Usando l'espressione (3.1) della media campionaria, si può scrivere *in primis*:

$$\begin{aligned} E\left[\overline{y}_{\mathbf{s}}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{N} y_i\, \delta(i;\, \mathbf{s})\right] \\ &= \frac{1}{n}\sum_{i=1}^{N} y_i\, E[\delta(i;\, \mathbf{s})]. \end{aligned} \tag{3.3}$$

In secondo luogo, si ha

$$\begin{aligned} E[\delta(i;\,\mathbf{s})] &= \sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s})\,\delta(i;\,\mathbf{s}) \\ &= \frac{1}{\binom{N}{n}} \sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\,\mathbf{s}). \end{aligned}$$

Nella $\sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\,\mathbf{s})$ si somma un 1 per ogni campione $\mathbf{s}$ contenente l'unità i (solo per questi campioni, infatti, è $\delta(i;\,\mathbf{s}) = 1$). Essa è perciò uguale al numero di campioni contenenti l'unità i, ossia al numero di combinazioni senza ripetizioni contenenti i, che è pari a $\binom{N-1}{n-1}$. Quindi:

$$E[\delta(i;\,\mathbf{s})] = \frac{\binom{N-1}{n-1}}{\binom{n}{N}} = \frac{n}{N}. \tag{3.4}$$

Inserendo la (3.4) nella (3.3) si ottiene

$$\begin{aligned} E\left[\overline{y}_{\mathbf{s}}\right] &= \frac{1}{n} \sum_{i=1}^{N} y_i \frac{n}{N} \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \mu_y \end{aligned}$$

ossia la (3.2). □

Un'osservazione importante: poiché $\delta(i;\,\mathbf{s})$ assume solo i valori 1 o 0 a seconda che il campione $\mathbf{s}$ contenga o meno l'unità i, il suo valore atteso è pari a

$$E[\delta(i;\,\mathbf{s})] = 0 \times Pr(\delta(i;\,\mathbf{s}) = 0) + 1 \times Pr(\delta(i;\,\mathbf{s}) = 1) = Pr(\delta(i;\,\mathbf{s}) = 1) = \frac{n}{N}.$$

Esso rappresenta la probabilità di selezionare un campione contenente l'unità i. Come si vedrà nel Capitolo 12, si tratta della *probabilità di inclusione* dell'unità i.

Essendo $\overline{y}_{\mathbf{s}}$ corretto, il suo errore quadratico medio coincide con la sua varianza: $MSE(\overline{y}_{\mathbf{s}}) = V(\overline{y}_{\mathbf{s}})$. Il calcolo di quest'ultima, non difficile, è svolto nella successiva proposizione.

Proposizione 3.2. *Se il disegno campionario è ssr, la varianza della media campionaria è pari a:*

$$V\left(\overline{y}_{\mathbf{s}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2\,. \tag{3.5}$$

Dimostrazione. Usando ancora l'espressione (3.1), si ha anzitutto:

$$
\begin{aligned}
V(\overline{y}_{\mathbf{s}}) &= E\left[\left(\frac{1}{n}\sum_{i\in\mathbf{s}} y_i - \mu_y\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i\in\mathbf{s}} (y_i - \mu_y)\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{N} (y_i - \mu_y)\,\delta(i;\,\mathbf{s})\right)^2\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N} (y_i - \mu_y)\,(y_j - \mu_y)\,\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{N} (y_i - \mu_y)^2\,\delta(i;\,\mathbf{s})^2\right. \\
&\qquad \left. + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} (y_i - \mu_y)\,(y_j - \mu_y)\,\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N} (y_i - \mu_y)^2\,E[\delta(i;\,\mathbf{s})] \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} (y_i - \mu_y)\,(y_j - \mu_y)\,E[\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})] \qquad (3.6)
\end{aligned}
$$

in quanto $\delta(i;\,\mathbf{s})^2 = \delta(i;\,\mathbf{s})$. Si è già visto che $E[\delta(i;\,\mathbf{s})] = n/N$. Inoltre, per ogni coppia i, j di unità *distinte* si ha

$$
\begin{aligned}
E[\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})] &= \sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s})\,\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s}) \\
&= \frac{1}{\binom{N}{n}}\sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s}).
\end{aligned}
$$

Nella

$$
\sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})
$$

si somma un 1 per ogni campione $\mathbf{s}$ contenente sia l'unità i che l'unità j, in quanto soltanto per questi campioni il prodotto $\delta(i;\,\mathbf{s})\delta(j;\,\mathbf{s})$ è uguale a 1. Pertanto, la somma in questione è uguale al numero di combinazioni senza

ripetizioni contenenti sia i che j, che è noto essere uguale a $\binom{N-2}{n-2}$. Ne consegue che:

$$E[\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s})] = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}. \tag{3.7}$$

Usando la (3.7) in (3.6) si ottiene infine

$$\begin{aligned}
V\left(\overline{y}_{\mathbf{s}}\right) &= \frac{1}{n^2}\sum_{i=1}^{N}(y_i-\mu_y)^2\,\frac{n}{N} + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(y_i-\mu_y)\,(y_j-\mu_y)\,\frac{n(n-1)}{N(N-1)} \\
&= \frac{1}{nN}\sum_{i=1}^{N}(y_i-\mu_y)^2 + \frac{1}{n}\frac{n-1}{N(N-1)}\sum_{i=1}^{N}(y_i-\mu_y)\sum_{\substack{j=1\\ j\neq i}}^{N}(y_j-\mu_y) \\
&= \frac{1}{n}\sigma_y^2 + \frac{1}{n}\frac{n-1}{N(N-1)}\left(\sum_{i=1}^{N}(y_i-\mu_y)\left\{\sum_{j=1}^{N}(y_j-\mu_y)-(y_i-\mu_y)\right\}\right) \\
&= \frac{1}{n}\sigma_y^2 + \frac{1}{n}\frac{n-1}{N(N-1)}\left\{\sum_{i=1}^{N}(y_i-\mu_y)\sum_{j=1}^{N}(y_j-\mu_y) - \sum_{i=1}^{N}(y_i-\mu_y)^2\right\} \\
&= \frac{1}{n}\sigma_y^2 \\
&\quad + \frac{1}{n}\frac{n-1}{N-1}\left\{\sum_{i=1}^{N}(y_i-\mu_y)\left(\frac{1}{N}\sum_{j=1}^{N}(y_j-\mu_y)\right) - \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_y)^2\right\} \\
&= \frac{1}{n}\sigma_y^2 - \frac{1}{n}\frac{n-1}{N-1}\sigma_y^2 \\
&= \frac{1}{n}\frac{N-n}{N-1}\sigma_y^2 \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2.
\end{aligned}$$

□

La relazione (3.5) mostra che la varianza della media campionaria, nel disegno ssr, dipende *solo* dalla varianza della popolazione, dalla numerosità campionaria, e da quella della popolazione. Essa si può anche riscrivere come:

$$V(\overline{y}_{\mathbf{s}}) = \frac{1}{n}S_y^2(1-f) \tag{3.8}$$

dove $f = n/N$ è la *frazione sondata* (o *frazione di campionamento*), cioè la frazione della popolazione soggetta a osservazione campionaria. Nella gran parte dei casi pratici, quando f è piccolo, diciamo dell'ordine del 5% o meno,

di fatto il termine $1 - f$ (detto *fattore di correzione per popolazione finita*) è virtualmente trascurabile nella formula (3.8). In tali casi, la numerosità della popolazione non ha praticamente nessuna influenza sulla varianza di $\overline{y}_{\mathbf{s}}$. Ad esempio se da una popolazione di 10000 unità si estrae un campione ssr di numerosità 100, si ha $f = 100/10000 = 0.01$. La media campionaria ha varianza

$$V(\overline{y}_{\mathbf{s}}) = \frac{1}{100} S_y^2 (1 - 0.01) = 0.0099 S_y^2. \tag{3.9}$$

Se invece un campione ssr di ampiezza 100 viene estratto da una popolazione di 10000000 di unità, a parità di valore di S_y^2 si ha

$$V(\overline{y}_{\mathbf{s}}) = \frac{1}{100} S_y^2 (1 - 0.00001) = 0.01 S_y^2. \tag{3.10}$$

Il rapporto tra (3.10) e (3.9) è pari a 1.01, per cui, a parità di varianza S_y^2, campionare 100 unità da una popolazione di 10000000 di unità anziché da una di 10000 unità fa aumentare la varianza della media campionaria (all'incirca) dell'1%.

Se l'effetto della numerosità della popolazione su $V(\overline{y}_{\mathbf{s}})$ è, nella gran parte dei casi, pressoché trascurabile, lo stesso non si può dire degli altri due termini, S_y^2 e n. In particolare, fermi restando N e S_y^2, sui quali non si ha alcuna influenza, dalla (3.5) si vede che al crescere della numerosità campionaria n la $V(\overline{y}_{\mathbf{s}})$ decresce alla velocità di $1/n$. In altri termini, $V(\overline{y}_{\mathbf{s}})$ *è dell'ordine di* $1/n$.

La varianza dello stimatore $\overline{y}_{\mathbf{s}}$ esprime in termini *assoluti* la variabilità di $\overline{y}_{\mathbf{s}}$ intorno a μ_y. Si tratta quindi di una misura *assoluta* dell'imprecisione di $\overline{y}_{\mathbf{s}}$ nello stimare μ_y. In molti casi è di interesse disporre anche di una misura *relativa*, che esprima l'imprecisione di $\overline{y}_{\mathbf{s}}$ in termini della media μ_y da stimare. Ad esempio, si supponga che la varianza della media campionaria sia $V(\overline{y}_{\mathbf{s}}) = 100$, così che $\sqrt{V(\overline{y}_{\mathbf{s}})} = 10$. Se si dovesse stimare una media $\mu_y = 1000$, $\overline{y}_{\mathbf{s}}$ verrà giudicato uno stimatore "preciso", perché l'errore di stima che in media si commetterebbe sarebbe dell'ordine della centesima parte di μ_y, e quindi "piccolo" rispetto a μ_y. Ma se fosse $\mu_y = 1$, la media campionaria $\overline{y}_{\mathbf{s}}$ sarebbe uno stimatore estremamente impreciso, in quanto l'errore di stima sarebbe in media dieci volte più grande di μ_y.

La più semplice misura relativa è il *coefficiente di variazione* di $\overline{y}_{\mathbf{s}}$, pari a

$$CV(\overline{y}_{\mathbf{s}}) = \frac{\sqrt{V(\overline{y}_{\mathbf{s}})}}{|\mu_y|} 100 = \sqrt{\frac{1}{n} - \frac{1}{N}} \frac{S_y}{|\mu_y|} 100. \tag{3.11}$$

3.3 Stima della varianza

Benché gran parte della nostra trattazione si concentri sulla stima (con diversi disegni campionari) della media di una popolazione, questa non è il solo parametro di interesse in rilevazioni campionarie. In diversi casi è anche di interesse

cercare di ottenere informazioni sulla variabilità della popolazione oggetto di studio. Poiché il principale indice di variabilità è la varianza, ci si concentrerà esclusivamente sul problema della sua stima sulla base dei dati campionari. I risultati ottenuti fino ad ora permettono in effetti di costruire facilmente uno stimatore della varianza corretta della popolazione, S_y^2. Definiamo *varianza campionaria corretta* la quantità

$$\widehat{s}_y^2 = \frac{1}{n-1}\sum_{i\in\mathbf{s}}(y_i - \overline{y}_\mathbf{s})^2. \tag{3.12}$$

Il primo risultato della presente sezione è che la (3.12) è uno stimatore non distorto di S_y^2. Più avanti, verrà affrontato il problema di stimare la varianza della media campionaria, $V(\overline{y}_\mathbf{s})$.

Proposizione 3.3. *Se il disegno campionario è ssr, la varianza campionaria corretta è uno stimatore corretto della varianza corretta della popolazione:*

$$E[\widehat{s}_y^2] = S_y^2. \tag{3.13}$$

Dimostrazione. In primo luogo, osservando che

$$\begin{aligned}
\sum_{i\in\mathbf{s}}(y_i - \overline{y}_\mathbf{s})^2 &= \sum_{i\in\mathbf{s}}\{(y_i - \mu_y) - (\overline{y}_\mathbf{s} - \mu_y)\}^2 \\
&= \sum_{i\in\mathbf{s}}\left\{(y_i - \mu_y)^2 + (\overline{y}_\mathbf{s} - \mu_y)^2 - 2(\overline{y}_\mathbf{s} - \mu_y)(y_i - \mu_y)\right\} \\
&= \sum_{i\in\mathbf{s}}(y_i - \mu_y)^2 + n(\overline{y}_\mathbf{s} - \mu_y)^2 - 2(\overline{y}_\mathbf{s} - \mu_y)\sum_{i\in\mathbf{s}}(y_i - \mu_y) \\
&= \sum_{i\in\mathbf{s}}(y_i - \mu_y)^2 - n(\overline{y}_\mathbf{s} - \mu_y)^2
\end{aligned}$$

$\widehat{s}_y^2$ si può riscrivere come

$$\widehat{s}_y^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i\in\mathbf{s}}(y_i - \mu_y)^2 - (\overline{y}_\mathbf{s} - \mu_y)^2\right). \tag{3.14}$$

Usando gli stessi calcoli della Proposizione 3.2, si ha poi

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i\in\mathbf{s}}(y_i - \mu_y)^2\right] &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \\
&= \frac{N-1}{N}S_y^2.
\end{aligned} \tag{3.15}$$

Inoltre, è evidente che

$$E\left[(\overline{y}_\mathbf{s} - \mu_y)^2\right] = V(\overline{y}_\mathbf{s}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2$$

per cui dalle (3.14), (3.15) si ottiene

$$\begin{aligned} E[\widehat{s}_y^2] &= \frac{n}{n-1}\left\{\left(1-\frac{1}{N}\right)S_y^2 - \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2\right\} \\ &= \frac{n}{n-1}\left(1-\frac{1}{n}\right)S_y^2 \\ &= S_y^2 \end{aligned}$$

il che completa la dimostrazione della (3.13). □

Naturalmente, dalla Proposizione 3.3 discende subito che

$$\widehat{\sigma}_y^2 = \frac{N-1}{N}\widehat{s}_y^2 = \frac{N-1}{N}\frac{1}{n-1}\sum_{i\in\mathbf{s}}(y_i-\overline{y}_\mathbf{s})^2$$

è uno stimatore corretto della varianza σ_y^2 della popolazione. Il termine $(N-1)/N$ è essenzialmente un fattore correttivo dovuto al fatto che la popolazione di riferimento è finita, di numerosità N . Naturalmente, a meno che N non sia piccolo, si ha $(N-1)/N \approx 1$, da cui $\widehat{\sigma}_y^2 \approx \widehat{s}_y^2$.

Come sottoprodotto della Proposizione 3.3 si ottiene uno stimatore corretto della varianza della media campionaria. È appena il caso di sottolineare l'importanza di questo risultato, in quanto, essendo $\overline{y}_\mathbf{s}$ corretto (quando usato con il disegno ssr), quello che si ottiene è uno stimatore corretto del suo errore quadratico medio (l'importanza di questo fatto è sottolineata nella Sezione 2.4).

Proposizione 3.4. *Se il disegno campionario è ssr, la quantità:*

$$\widehat{V} = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in\mathbf{s}}(y_i-\overline{y}_\mathbf{s})^2 \tag{3.16}$$

è uno stimatore corretto di $V(\overline{y}_\mathbf{s})$.

Dimostrazione. È un'immediata conseguenza di (3.5) e (3.13). □

I risultati finora ottenuti ci permettono anche di stimare il coefficiente di variazione della media campionaria. Tenendo conto dell'espressione (3.11), uno stimatore di $CV(\overline{y}_\mathbf{s})$ è il seguente:

$$\widehat{CV} = \frac{\sqrt{\widehat{V}}}{|\overline{y}_\mathbf{s}|}\,100. \tag{3.17}$$

Come si vede dalla (3.17), sia il numeratore che il denominatore di $\widehat{CV}$ dipendono dal campione $\mathbf{s}$, quindi in generale variano al variare del campione stesso. Questo fa sì che il calcolo del valore atteso di $\widehat{CV}$ non sia così semplice come quelli visti finora. In generale, non è difficile verificare che $\widehat{CV}$ è uno stimatore *distorto* del coefficiente di variazione $CV(\overline{y}_\mathbf{s})$.

3.4 Approssimazione normale nel disegno ssr e intervalli di confidenza per la media della popolazione

Se si utilizza l'approssimazione normale si possono anche costruire intervalli di confidenza approssimati per la media della popolazione. L'idea di base, già accennata nella Sezione 2.6, è quella di approssimare l'effettiva distribuzione di probabilità della media campionaria *studentizzata*

$$Z = \frac{\overline{y}_{\mathbf{s}} - \mu_y}{\sqrt{\widehat{V}}} \tag{3.18}$$

con una normale standard $N(0, 1)$. Se $z_{\alpha/2}$ è il valore tale che $Pr(N(0, 1) \geqslant z_{\alpha/2}) = \alpha/2$, si ha in via approssimata

$$Pr\left(-z_{\alpha/2} \leqslant \frac{\overline{y}_{\mathbf{s}} - \mu_y}{\sqrt{\widehat{V}}} \leqslant z_{\alpha/2}\right) = 1 - \alpha$$

ossia

$$Pr\left(\overline{y}_{\mathbf{s}} - z_{\alpha/2}\sqrt{\widehat{V}} \leqslant \mu_y \leqslant \overline{y}_{\mathbf{s}} + z_{\alpha/2}\sqrt{\widehat{V}}\right) = 1 - \alpha.$$

Ne consegue che l'intervallo

$$\left[\overline{y}_{\mathbf{s}} - z_{\alpha/2}\sqrt{\widehat{V}},\ \overline{y}_{\mathbf{s}} + z_{\alpha/2}\sqrt{\widehat{V}}\right] \tag{3.19}$$

è un intervallo di confidenza approssimato per la media μ_y della popolazione al livello $1 - \alpha$.

Esempio 3.3. Nel *file* `stature.txt` sono riportati numeri di matricola, sesso e statura di una popolazione fittizia di $N = 1570$ studenti universitari. I numeri di matricola sono inventati, e gli altri dati sono generati mediante simulazione. La statura media della popolazione è $\mu_y = 172.8$, e la varianza è pari a $\sigma_y^2 = 59.9$. Da questa popolazione bisogna selezionare un campione di $n = 50$ unità, e costruire un intervallo di confidenza, al livello 0.95, per la media della popolazione. Immaginiamo di etichettare le unità con numeri interi, da 1 a 1570. Le etichette delle unità campionarie sono qui sotto riportate:

992 1062 1265 1112 487 987 1289 1170 1296 942 97 329 1391 455 311 906
1403 661 1090 1127 1417 537 632 662 1400 11 347 850 275 1361 178 291 1385
916 695 723 965 1272 1521 905 584 399 1288 238 159 561 1064 1465 71 973.

I corrispondenti dati campionari sono riportati nel *file* `campstature.txt`. La media campionaria è $\overline{y}_{\mathbf{s}} = 172.76$, e la varianza campionaria corretta risulta eguale a $\widehat{s}_y^2 = 34.06$. Lo stimatore $\widehat{V}$ (3.16) assume in questo caso il valore 0.66. Tenendo infine conto che $z_{0.025} = 1.96$, si conclude che l'intervallo

$$\left[172.76 - 1.96\sqrt{0.66},\ 172.76 + 1.96\sqrt{0.66}\right] = [171.17,\ 174.35]$$

è un intervallo di confidenza per μ_y al livello (approssimato) 0.95. □

La validità dell'intervallo di confidenza (3.19) merita qualche precisazione. Il suo fondamento *teorico* poggia sul teorema limite centrale per il campionamento ssr da popolazioni finite; l'articolo-chiave in questa direzione è quello di Hájek (1960). Le condizioni di regolarità sono non difficili, ma piuttosto "astratte", in quanto richiedono che sia la numerosità della popolazione che quella del campione tendano all'infinito, seppure con velocità diversa. Immaginiamo di avere una *successione* di popolazioni I_{N_ν}, $\nu = 1, 2, \ldots$, di numerosità N_ν, $\nu = 1, 2, \ldots$. Indichiamo $y_{i\nu}$ la modalità dell'unità $i$ma della popolazione I_{N_ν}, $i = 1, \ldots, N_\nu$, $\nu = 1, 2, \ldots$, e siano

$$\mu_{y\nu} = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} y_{i\nu}, \quad S^2_{y\nu} = \frac{1}{N_\nu - 1} \sum_{i=1}^{N_\nu} (y_{i\nu} - \mu_{y\nu})^2$$

rispettivamente la media e la varianza corretta della popolazione I_{N_ν}.

Dalla popolazione I_{N_ν} si seleziona, mediante campionamento ssr, un campione di n_ν unità, $\nu = 1, 2, \ldots$. Indichiamo con $\overline{y}_\nu$ la media campionaria. Per ϵ positivo, sia poi $E_{\nu\epsilon}$ l'insieme delle unità di I_{N_ν} tali che

$$|y_{i\nu} - \mu_{y\nu}| > \epsilon \sqrt{n_\nu (1 - f_\nu)} S_\nu$$

con $f_\nu = n_\nu / N_\nu$. Si può dimostrare che se:

- N_ν e n_ν tendono all'infinito al crescere di ν;
- $N_\nu - n_\nu$ tende all'infinito al crescere di ν;
- $\lim_{\nu \to \infty} \left(\frac{1}{N\nu} 8 \sum_{i \in E_{\nu\epsilon}} (y_{i\nu} - \mu_{y\nu})^2 \right) / S^2_{y\nu} = 0$ per ogni $\epsilon > 0$

allora

$$\lim_{\nu \to \infty} Pr \left(\frac{\overline{y}_\nu - \mu_{y\nu}}{\sqrt{V(\overline{y}_\nu)}} \leqslant z \right) = Pr(N(0,\, 1) \leqslant z) \;\; \textit{per ogni } z \textit{ reale.} \qquad (3.20)$$

Si può anche provare che, sotto condizioni aggiuntive (la più importante delle quali è che esistano i limiti $\lim_{\nu \to \infty} \mu_{y\nu}$, $\lim_{\nu \to \infty} S^2_{y\nu}$, e che il secondo limite sia positivo), il limite (3.20) continua a valere se la varianza della media campionaria $V(\overline{y}_\nu)$ è sostituita con la sua stima campionaria $\widehat{V}$.

In pratica, l'approssimazione normale su cui si basa l'intervallo di confidenza (3.19) è valida purché sia la numerosità campionaria n, sia la differenza $N - n$ tra numerosità della popolazione e numerosità del campione, siano "sufficientemente grandi". Ciò ha luogo quando la numerostà del campione da un lato è "grande", ma dall'altro è "abbastanza piccola" rispetto alla numerosità della popolazione. Spesso viene fornito il numero "magico" $n = 30$ come valore di soglia per la numerosità campionaria, a partire dal quale è lecito usare l'approssimazione normale. Per la verità, non è così semplice rispondere alla domanda: "*A partire da quale numerosità campionaria è valida l'approssimazione normale*?" Molto dipende dall'asimmetria della popolazione. Se la popolazione è vicina alla simmetria, una numerosità campionaria $n = 30$ è probabilmente bastevole per usare l'approssimazione normale. Se però la popolazione è fortemente asimmetrica, è necessario usare una numerosità campionaria molto maggiore. I successivi esempi chiariscono questo punto.

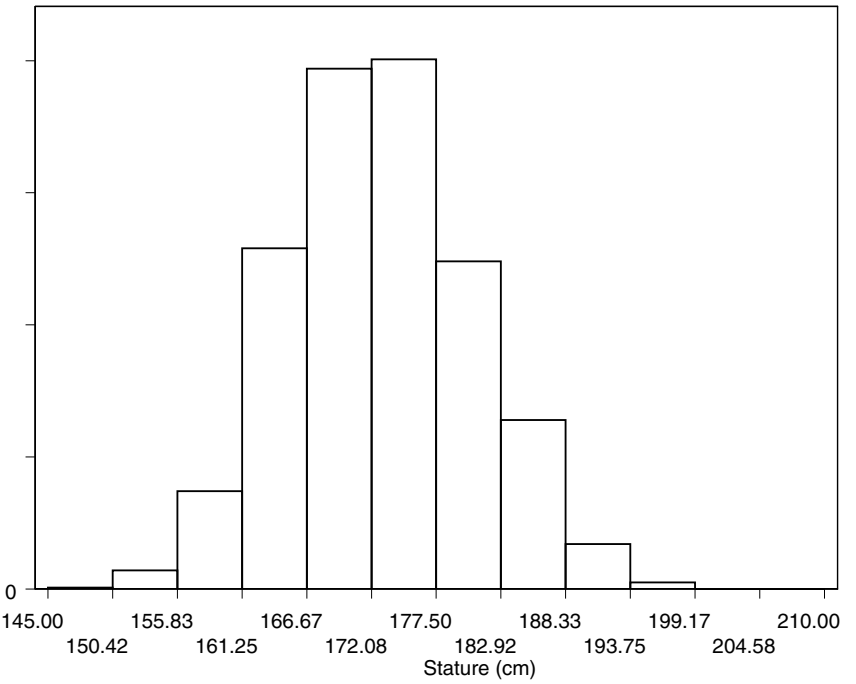


Fig. 3.1 Istogramma delle stature di una popolazione di 1570 studenti

Esempio 3.4. Consideriamo la popolazione di 1570 studenti del *file* `stature.txt` dell'Esempio 3.3. In Fig. 3.1 è riportato l'istogramma delle stature. L'indice di asimmetria

$$\gamma_y = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \mu_y}{\sigma_y}\right)^3 \tag{3.21}$$

risulta pari a 0.18. Si tratta quindi di una popolazione solo moderatamente asimmetrica, molto vicina a una situazione di simmetria, fatto confermato anche visivamente dalla Fig. 3.1.

Per questa popolazione si è considerato un disegno ssr di numerosità $n = 30$. Per studiare la distribuzione di probabilità della media studentizzata (3.18), sono stati generati, sempre mediante simulazione, 1000 campioni ssr indipendenti, per ognuno dei quali è stato calcolato il rapporto (3.18). Il relativo istogramma, mostrato in Fig. 3.2, fornisce una buona approssimazione della distribuzione di probabilità della (3.18). Chiaramente, in questo caso l'uso dell'approssimazione normale per la distribuzione di probabilità della media studentizzata (3.18) è perfettamente valido. □

Esempio 3.5. Nel *file* `cultura.txt` sono riportate le spese annue per attività culturali di 1500 famiglie. Si tratta di dati ottenuti modificando opportunamente dati reali rilevati, nel corso di vari anni, da studenti della Facoltà di

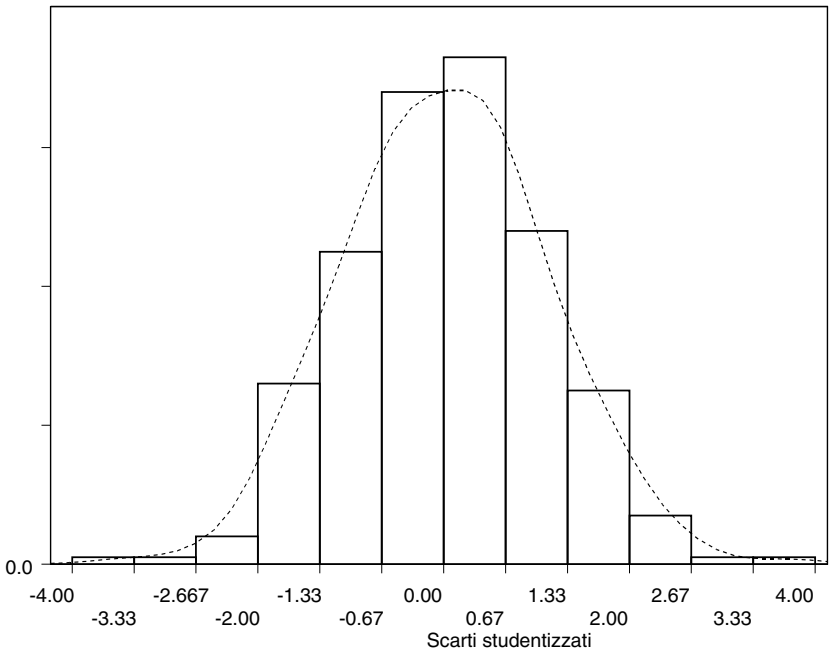


Fig. 3.2 Distribuzione di probabilità della media campionaria studentizzata (n=30) per la popolazione di Fig. 3.1

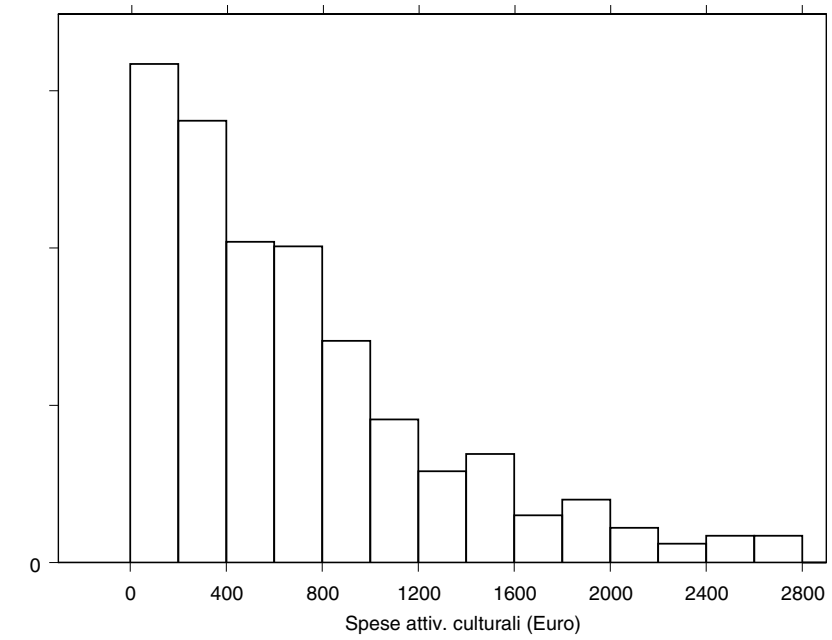


Fig. 3.3 Istogramma delle spese per attività culturali di 1500 famiglie

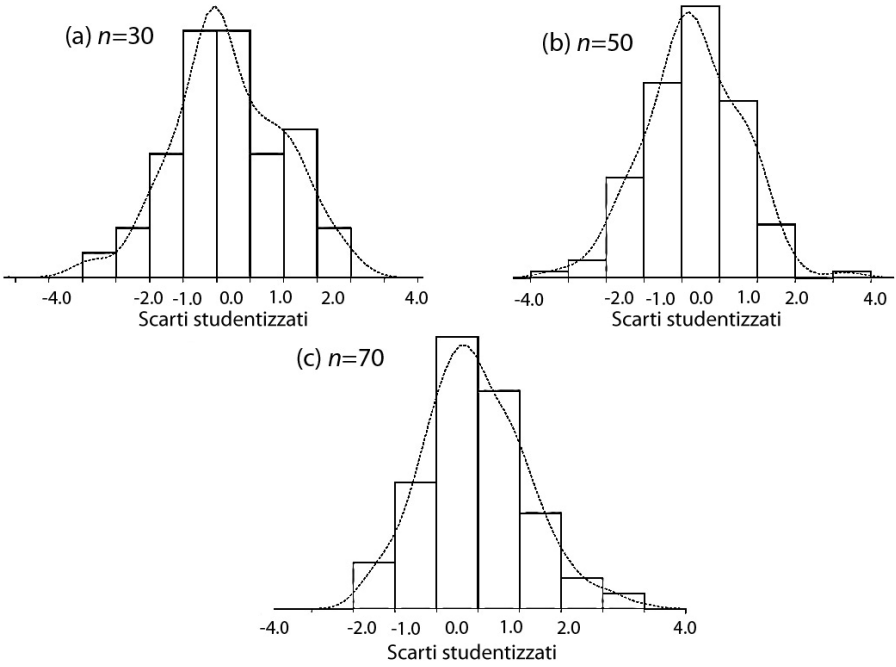


Fig. 3.4 Distribuzione di probabilità della media campionaria studentizzata (n=30, 50, 70) per la popolazione di Fig. 3.3

Scienze Statistiche dell'Università di Roma "La Sapienza". In Fig. 3.3 è riportato l'istogramma delle spese per attività culturali. Questo carattere, evidentemente, ha una distribuzione molto asimmetrica. La media della popolazione è $\mu_y = 702.5$, e la deviazione *standard* è $\sigma_y = 592.6$. Si tratta di una popolazione con alta variabilità (il coefficiente di variazione è del 84%, con un grande gruppo di famiglie che ha spese relativamente contenute, ed un gruppo non trascurabile con spese alte. L'indice di asimmetria γ (3.21) assume un valore 1.3, il che evidenzia la forte asimmetria positiva della popolazione.

La Fig. 3.4 raffigura la distribuzione di probabilità (ottenuta simulando 1000 campioni indipendenti) della media campionaria studentizzata, rispettivamente per campioni di numerosità (a) $n = 30$, (b) $n = 50$, (c) $n = 70$.

La presenza di una forte asimmetria peggiora, rispetto al caso dell'esempio precedente, l'approssimazione normale. In particolare, questa diviene accettabile solo per una numerosità campionaria $n = 70$. □

3.5 Un importante caso speciale: la stima di proporzioni

La stima di proporzioni può essere trattata come caso particolare di stima della media di una popolazione. Supponiamo di essere interessati a stimare la proporzione di unità della popolazione che possiedono un determinato attributo, diciamo A. Come già fatto nell'Es. 1.1, poniamo in questo caso:

$$y_i = \begin{cases} 1 \ \textit{se l'unità } i \textit{ possiede l'attributo } A \\ 0 \ \textit{altrimenti} \end{cases} \qquad \textit{per ciascuna unità } i = 1, \ldots, N.$$

Indichiamo con N_A il numero di unità della popolazione che possiedono l'attributo A, e con P_A la proporzione di unità della popolazione che possiedono A. Come già visto, dalla relazione $N_A = \sum_{i=1}^{N} y_i$ discende che

$$P_A = \frac{N_A}{N} = \frac{1}{N} \sum_{i=1}^{N} y_i = \mu_y.$$

La varianza della popolazione, inoltre, è pari a

$$\sigma_y^2 = P_A(1 - P_A).$$

Se indichiamo ora con n_A il numero di unità del campione che possiedono l'attributo A, e con $\widehat{p}_A = n_A/n$ la proporzione di unità campionarie che possiedono A, vale l'ovvia relazione

$$n_A = \sum_{i \in \mathbf{s}} y_i$$

dalla quale discende che:

$$\widehat{p}_A = \frac{n_A}{n} = \frac{1}{n} \sum_{i \in \mathbf{s}} y_i = \overline{y}_{\mathbf{s}}$$

ovvero *la media campionaria coincide con la proporzione di unità campionarie che possiedono l'attributo A.*

A questo punto è facile particolarizzare i risultati ottenuti nelle Sezioni 3.2 e 3.3. In primo luogo, $\widehat{p}_A$ è uno stimatore corretto di P_A:

$$E[\widehat{p}_A] = P_A.$$

Tenendo poi conto che $S_y^2 = \frac{N}{N-1}\sigma_y^2$, dalla (3.5) si ottiene la relazione:

$$\begin{aligned} V(\widehat{p}_A) &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N}{N-1} P_A(1 - P_A) \\ &= \frac{N-n}{N-1} \frac{P_A(1-P_A)}{n}. \end{aligned} \tag{3.22}$$

Naturalmente, a meno che N non sia molto piccolo si ha $N - 1 \approx N$, per cui la (3.22) si riduce a

$$V(\widehat{p}_A) \approx \left(\frac{1}{n} - \frac{1}{N}\right) P_A(1 - P_A).$$

L'uguaglianza (3.22) evidenzia che per $P_A = 0$ o $P_A = 1$ si ha $V(\widehat{p}_A) = 0$. In sostanza questo significa che $V(\widehat{p}_A)$ assume valori molto piccoli nei casi estremi in cui la proporzione P_A da stimare assume valori prossimi a 0 o a 1. Apparentemente, questo ragionamento sembrerebbe suggerire che l'uso della proporzione campionaria $\widehat{p}_A$ per stimare P_A fornisce risultati molto precisi quando quest'ultima è o molto piccola, o molto grande. Le cose stanno però in modo diverso se si valuta l'errore di stima in termini *relativi*, ossia se si fa riferimento al coefficiente di variazione di $\widehat{p}_A$. Questo risulta pari a:

$$\begin{aligned} CV(\widehat{p}_A) &= \frac{\sqrt{V(\widehat{p}_A)}}{E[\widehat{p}_A]} \\ &= \sqrt{\frac{N-n}{n(N-1)}} \frac{\sqrt{P_A(1-P_A)}}{P_A} \\ &= \sqrt{\frac{N-n}{n(N-1)}} \sqrt{\frac{1}{P_A} - 1}. \end{aligned} \tag{3.23}$$

La (3.23) evidenzia che:

1. se P_A cresce verso 1, $CV(\widehat{p}_A)$ tende a 0;
2. se P_A decresce verso 0, $CV(\widehat{p}_A)$ tende all'infinito.

L'asserzione 1 mette in evidenza che la proporzione campionaria $\widehat{p}_A$ dà risultati molto precisi, non solo in termini assoluti ma anche relativi, quando la proporzione P_A da stimare è molto grande. Invece, sempre in termini relativi, $\widehat{p}_A$ fornisce cattivi risultati quando la proporzione P_A da stimare è piccola, prossima a zero. Intuitivamente, la stima di una proporzione P_A "piccola" è un problema assai difficile in quanto, a meno che la numerosità campionaria n non sia molto alta, con elevata probabilità si osservano nel campione solo pochissime unità che possiedono l'attributo A.

La Proposizione 3.4 suggerisce uno stimatore corretto di $V(\widehat{p}_A)$. Tenendo conto (come facilmente si vede) che

$$\frac{1}{n} \sum_{i \in \mathbf{s}} (y_i - \overline{y}_{\mathbf{s}})^2 = \widehat{p}_A(1 - \widehat{p}_A)$$

dalla (3.16) si ha che

$$\begin{aligned} \widehat{V} &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in \mathbf{s}} (y_i - \overline{y}_{\mathbf{s}})^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{n-1} \widehat{p}_A(1 - \widehat{p}_A) \end{aligned} \tag{3.24}$$

è uno stimatore corretto di $V(\widehat{p}_A)$. Naturalmente, a meno che la numerosità campionaria n non sia piccola, si ha $n - 1 \approx n$, per cui per lo stimatore $\widehat{V}$ vale l'approssimazione

$$\widehat{V} \approx \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{p}_A(1 - \widehat{p}_A). \tag{3.25}$$

Dai risultati della Sezione 3.4, infine, è possibile costruire un intervallo di confidenza approssimato per P_A. Particolarizzando infatti la (3.19), e lasciando per il resto invariata la notazione già usata, si ha che l'intervallo

$$\left[\widehat{p}_A - z_{\alpha/2}\sqrt{\widehat{V}},\ \widehat{p}_A + z_{\alpha/2}\sqrt{\widehat{V}}\right] \tag{3.26}$$

è un intervallo di confidenza approssimato per P_A al livello $1 - \alpha$, con $\widehat{V}$ dato dalla (3.24). Per quanto riguarda l'accuratezza dell'intervallo di confidenza (3.26), valgono considerazioni abbastanza simili a quelle svolte nella Sezione 3.4. Per approfondimenti e ulteriori considerazioni, si rinvia al volume di Cochran (1977) (pp. 57-59).

Esempio 3.6. Con riferimento ai dati contenuti nel *file* `cultura.txt` (cfr. Esempio 3.5), si vuole stimare, sulla base di un campione ssr di numerosità $n = 100$, la proporzione di individui che spendono ogni anno per cultura più di 1000 Euro.

Per ognuna delle 1500 unità della popolazione, poniamo $y_i = 1$ se l'unità i spende più di 1000 Euro, e $y_i = 0$ in caso contrario ($i = 1, \ldots, 1000$). Nel *file* `spese+1000.txt` sono riportati i dati per l'intera popolazione. Il numero totale di unità della popolazione che spendono annualmente più di 1000 Euro è $N_A = 356$; la proporzione di unità della popolazione che spendono più di 1000 Euro è $P_A = 0.237$. La varianza è pari a $P_A(1 - P_A) = 0.181$.

Dalla popolazione è selezionato, come detto all'inizio, un campione di $n = 100$ unità. L'elenco delle unità campionarie, con i relativi dati osservati, è riportato nel *file* `campione1.txt`. Il numero di unità del campione che spendono ogni anno più di 1000 Euro è $n_A = 22$, così che la corrispondente proporzione campionaria è $\widehat{p}_A = 22/100 = 0.22$. La stima della varianza di $\widehat{p}_A$, ottenuta in base alla (3.24), è $\widehat{V} = 0.00185$. Pertanto, un intervallo di confidenza per P_A, al livello approssimato 0.95, è il seguente:

$$\left[0.22 - 1.96\sqrt{0.0018},\ 0.22 + 1.96\sqrt{0.0018}\right] = [0.137, 0.303].$$

Se si usa la formula approssimata (3.25), si ottiene una stima $\widehat{V} = 0.00183$, e quindi un intervallo di confidenza per P_A praticamente uguale al precedente. □

3.6 Regola di estensione per la stima di parametri lineari

I risultati ottenuti per la stima della media della popolazione valgono per parametri di interesse molto più generali: i *parametri lineari*. Un parametro $\theta = \theta(\boldsymbol{Y}_N)$ è un parametro lineare se per ogni unità i della popolazione è definita una funzione $t_i(\cdot)$ (a valori reali) tale che si può scrivere:

$$\theta = \frac{1}{N}\sum_{i=1}^{N} t_i(y_i). \tag{3.27}$$

Le N funzioni $t_1(\cdot), \ldots, t_N(\cdot)$ non hanno tutte necessariamente la stessa forma.

Ovviamente, la media campionaria è una caso speciale di parametro lineare, con $t_i(y_i) = y_i$ per ogni $i = 1, \ldots, N$. Qui di seguito sono riportati alcuni altri semplici esempi di parametri lineari.

- *Ammontare del carattere* $\mathcal{Y}$ nella popolazione. Se si pone $t_i(y_i) = Ny_i$, si ha $\theta = \frac{1}{N}\sum_{i=1}^{N} Ny_i = \sum_{i=1}^{N} y_i$, che è l'ammontare del carattere $\mathcal{Y}$ nella popolazione di riferimento.
- *Momento* $k$mo (dall'origine) del carattere $\mathcal{Y}$. Se si pone $t_i(y_i) = y_i^k$, si ha $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i^k$ = momento $k$mo di $\mathcal{Y}$.
- *Funzione di ripartizione* del carattere $\mathcal{Y}$. Per ogni fissato y reale, la funzione di ripartizione di $\mathcal{Y}$ nel punto y è definita come:

$$F(y) = \frac{\#\ \textit{di unità della popolazione tali che}\ y_i \leqslant y}{N}.$$

Anche $F(y)$ si può esprimere come parametro lineare. Poniamo infatti

$$t_i(y_i) = \begin{cases} 1\ \textit{se}\ y_i \leqslant y \\ 0\ \textit{se}\ y_i > y \end{cases} \quad \textit{per ciascuna unità}\ i = 1, \ldots, N.$$

La $\sum_{i=1}^{N} t_i(y_i)$ somma tanti 1 quante sono le unità i tali che $y_i \leqslant y$, e tanti 0 quante sono le unità i tali che $y_i > y$. Ne consegue che:

$$\sum_{i=1}^{N} t_i(y_i) = \#\ \textit{di unità della popolazione tali che}\ y_i \leqslant y$$

per cui si può scrivere:

$$\theta = \frac{1}{N}\sum_{i=1}^{N} t_i(y_i) = F(y).$$

Ritornando agli aspetti generali riguardanti la stima di un parametro lineare, se per ogni unità i della popolazione si pone $z_i = t_i(y_i)$, resta definito

un nuovo carattere $\mathcal{Z}$, che assume le N modalità $z_1, \ldots, z_N$. La sua media, nella popolazione di riferimento, è pari a

$$\begin{aligned} \mu_z &= \frac{1}{N}\sum_{i=1}^{N} z_i \\ &= \frac{1}{N}\sum_{i=1}^{N} t_i(y_i) \\ &= \theta \end{aligned} \tag{3.28}$$

e la sua varianza corretta

$$\begin{aligned} S_z^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \mu_z)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}(t_i(y_i) - \theta)^2. \end{aligned} \tag{3.29}$$

La stima di θ, come mostrato dalla (3.28), può essere vista come un problema di stima di una media. Se si utilizza il disegno ssr, un suo stimatore corretto è la media campionaria delle z_i, che si può scrivere come:

$$\begin{aligned} \widehat{\theta} &= \overline{z}_{\mathbf{s}} \\ &= \frac{1}{n}\sum_{i\in\mathbf{s}} z_i \\ &= \frac{1}{n}\sum_{i\in\mathbf{s}} t_i(y_i). \end{aligned} \tag{3.30}$$

Usando i risultati ottenuti nella Sezione 3.2 e la (3.28), è immediato vedere che lo stimatore (3.30) è uno stimatore corretto di θ:

$$E[\widehat{\theta}] = E[\overline{z}_{\mathbf{s}}] = \mu_z = \theta. \tag{3.31}$$

Sempre da risultati noti, e tenendo conto della (3.29), la varianza di $\widehat{\theta}$ è pari a

$$\begin{aligned} V(\widehat{\theta}) &= \left(\frac{1}{n} - \frac{1}{N}\right) S_z^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\frac{1}{N-1}\sum_{i=1}^{N}(t_i(y_i) - \theta)^2\right\}. \end{aligned} \tag{3.32}$$

La regola di costruzione dello stimatore (3.30) è detta *regola di estensione*. Usando i risultati ottenuti nelle precedenti sezioni, è anche facile stimare la (3.32), e costruire un intervallo di confidenza approssimato per θ.

Esempio 3.7. Come già detto all'inizio di questa sezione, se $t_i(y_i) = Ny_i$ il parametro θ diviene l'ammontare del carattere $\mathcal{Y}$: $\theta = \sum_{i=1}^{N} y_i$. Lo stimatore corretto (3.30) assume in questo caso la forma:

$$\widehat{\theta} = \frac{1}{n} \sum_{i \in \mathbf{s}} N y_i = N \overline{y}_{\mathbf{s}}. \tag{3.33}$$

La varianza dello stimatore (3.33) può essere anche valutata direttamente, senza ricorrere alla formula generale (3.32). È infatti immediato verificare che

$$V(\widehat{\theta}) = V(N\overline{y}_{\mathbf{s}}) = N^2 V(\overline{y}_{\mathbf{s}}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2.$$

Naturalmente, in questo caso uno stimatore non distorto della varianza di $\widehat{\theta}$ è semplicemente:

$$\widehat{V}(\widehat{\theta}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \widehat{s}_y^2 \tag{3.34}$$

con $\widehat{s}_y^2$ dato dalla (3.12).

Si consideri, a titolo di esemplificazione numerica, la popolazione di 1500 famiglie del file `cultura.txt`, da cui si è selezionato un campione ssr di $n = 100$ unità. I dati campionari sono riportati nel file `campione1.txt`. Si vuole stimare l'ammontare delle spese annue sostenute dalle famiglie per attività culturali. La media campionaria del carattere "spese annue per attività culturali" è $\overline{y}_{\mathbf{s}} = 662.433$ (Euro). Essendo $N = 1500$, lo stimatore (3.33) assume il valore $\widehat{\theta} = 1500 \times 662.43 = 993649.5$ (Euro). Come stima della varianza di $\widehat{\theta}$, usando la (3.34) e tenendo conto che $\widehat{s}_y^2 = 324270.96$, si ha $\widehat{V}(\widehat{\theta}) = 6809690160$. Usando questi risultati, e tenendo conto che $z_{0.05} = 1.645$, si ricava facilmente che un intervallo di confidenza per θ al livello (approssimato) 0.90 è [851829, 1135470]. □

3.7 Popolazioni multivariate: stima di covarianze

I risultati ottenuti fino ad ora si estendono abbastanza facilmente anche al caso di popolazioni su cui si osservano due o più caratteri. Per semplicità ci limitiamo qui al caso di due soli caratteri, in quanto l'estensione a più di due caratteri è molto semplice.

Supponiamo che sulle unità della popolazione di riferimento siano definite le modalità di due caratteri, diciamo $\mathcal{X}$, $\mathcal{Y}$. La situazione è essenzialmente quella descritta nella Sezione 1.3. Indichiamo con (x_i, y_i) le modalità assunte rispettivamente da $\mathcal{X}$ e da $\mathcal{Y}$ in corrispondenza dell'unità $i$ma della popolazione. Indichiamo inoltre con μ_x, μ_y, S_x^2, S_y^2 le medie e le varianze (corrette) rispettivamente di $\mathcal{X}$, $\mathcal{Y}$. Se $\mathbf{s}$ è il campione di unità selezionate, i dati statistici campionari sono rappresentati dal campione di modalità (etichettate) $(\boldsymbol{x}(\mathbf{s}), \boldsymbol{y}(\mathbf{s})) = \{(x_i, y_i);\ i \in \mathbf{s}\}$.

La stima di parametri univariati, quali le medie μ_x, μ_y, non presenta difficoltà: basta applicare quanto svolto nelle Sezioni 3.2, 3.3. Molto più interessante, nel presente contesto, è cercare di stimare la *covarianza* tra i due caratteri in esame:

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)\,(y_i - \mu_y) = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \mu_x\, \mu_y.$$

Analogamente a quanto visto nella Sezione 3.3, otterremo questo risultato come sottoprodotto della covarianza tra medie campionarie . Precisamente, consideriamo le due medie campionarie:

$$\overline{x}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} x_i, \;\; \overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} y_i$$

e, analogamente alla Sezione 3.2, sia

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

la covarianza corretta tra $\mathcal{X}$ e $\mathcal{Y}$.

Conosciamo già valori attesi e varianze di $\overline{x}_{\mathbf{s}}$, $\overline{y}_{\mathbf{s}}$. Nella successiva proposizione è calcolata la covarianza tra $\overline{x}_{\mathbf{s}}$ e $\overline{y}_{\mathbf{s}}$.

Proposizione 3.5. *Se il disegno campionario è ssr, la covarianza tra* $\overline{x}_{\mathbf{s}}$ *e* $\overline{y}_{\mathbf{s}}$ *è pari a:*

$$C\left(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy}\,. \tag{3.35}$$

Dimostrazione. La tecnica è esattamente la stessa della Proposizione 3.2, per cui la dimostrazione verrà svolta solo per sommi capi. In primo luogo, essendo $E[\overline{x}_{\mathbf{s}}] = \mu_x$ e $E[\overline{y}_{\mathbf{s}}] = \mu_y$, si ha $C(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}}) = E[(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y)]$ e usando la (3.1) si può scrivere:

$$\begin{aligned}
&E[(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y)] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{N}(x_i - \mu_x)\,\delta(i;\,\mathbf{s})\right)\left(\frac{1}{n}\sum_{j=1}^{N}(y_j - \mu_y)\,\delta(j;\,\mathbf{s})\right)\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu_x)(y_j - \mu_y)\,\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s})\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)\,E[\delta(i;\,\mathbf{s})] \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}(x_i - \mu_x)(y_j - \mu_y)\,E[\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s})].
\end{aligned} \tag{3.36}$$

La (3.36) è esattamente della stessa forma della (3.6), ad esclusione del fatto che i termini del tipo $(y_i - \mu_y)^2$ e $(y_i - \mu_y)(y_j - \mu_y)$ sono rispettivamente sostituiti da termini $(x_i - \mu_x)(y_i - \mu_y)$ e $(x_i - \mu_x)(y_j - \mu_y)$. Ad ogni modo, gli stessi calcoli della Proposizione 3.2 portano alla (3.35). □

La covarianza tra medie campionarie (3.35) ha una struttura praticamente identica a quella della varianza della media campionaria (3.5). L'unica differenza è che il termine S_y^2 è ora sostituito da S_{xy}. Naturalmente, valgono ancora le considerazioni già fatte a proposito di $V(\overline{x}_{\mathbf{s}})$. In particolare, il valore di $C(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}})$ è principalmente determinato da n e da S_{xy}, mentre in genere la numerosità N della popolazione ha un'influenza pressoché trascurabile. Inoltre, dati N e S_{xy}, $C(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}})$ è dell'ordine di grandezza di $1/n$.

Procedendo in maniera simile a quanto visto nella Sezione 3.3, si è ora in grado di costruire uno stimatore corretto di S_{xy}. Analogamente alla (3.12), definiamo *covarianza campionaria corretta* la quantità

$$\widehat{s}_{xy} = \frac{1}{n-1} \sum_{i \in \mathbf{s}} (x_i - \overline{x}_{\mathbf{s}})(y_i - \overline{y}_{\mathbf{s}}). \tag{3.37}$$

Nella Proposizione 3.6 dimostriamo che (3.12) è uno stimatore non distorto di S_{xy}.

Proposizione 3.6. *Se il disegno campionario è ssr, la covarianza campionaria corretta è uno stimatore corretto della covarianza corretta della popolazione:*

$$E[\widehat{s}_{xy}] = S_{xy}. \tag{3.38}$$

Dimostrazione. In primo luogo, analogamente alla Proposizione 3.3 si ha

$$\begin{aligned}
&\sum_{i \in \mathbf{s}} (x_i - \overline{x}_{\mathbf{s}})(y_i - \overline{y}_{\mathbf{s}}) = \sum_{i \in \mathbf{s}} \{(x_i - \mu_x) - (\overline{x}_{\mathbf{s}} - \mu_x)\} \{(y_i - \mu_y) - (\overline{y}_{\mathbf{s}} - \mu_y)\} \\
&= \sum_{i \in \mathbf{s}} \{(x_i - \mu_x)(y_i - \mu_y) - (x_i - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y) - (y_i - \mu_y)(\overline{x}_{\mathbf{s}} - \mu_x) \\
&\quad + (\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y)\} \\
&= \sum_{i \in \mathbf{s}} (x_i - \mu_x)(y_i - \mu_y) - n(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y) - n(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y) \\
&\quad + n(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y) \\
&= \sum_{i \in \mathbf{s}} (x_i - \mu_x)(y_i - \mu_y) - n(\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y)
\end{aligned}$$

per cui $\widehat{s}_{xy}$ si può riscrivere come

$$\widehat{s}_{xy} = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i \in \mathbf{s}} (x_i - \mu_x)(y_i - \mu_y) - (\overline{x}_{\mathbf{s}} - \mu_x)(\overline{y}_{\mathbf{s}} - \mu_y) \right).$$

Usando gli stessi calcoli della Proposizione 3.2, e tenendo conto che $E[\delta(i;\,\mathbf{s})] = n/N$, si ha poi

$$E\left[\frac{1}{n}\sum_{i\in\mathbf{s}}(x_i-\mu_x)(y_i-\mu_y)\right] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y) = \frac{N-1}{N}S_{xy}.$$

Essendo infine

$$E\left[(\overline{x}_{\mathbf{s}}-\mu_x)(\overline{y}_{\mathbf{s}}-\mu_y)\right] = C(\overline{x}_{\mathbf{s}},\,\overline{y}_{\mathbf{s}}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{xy}$$

è immediato verificare, ripetendo *verbatim* gli stessi calcoli della Proposizione 3.3, che vale la (3.38). □

Dalla Proposizione 3.6 si trae che

$$\widehat{\sigma}_{xy} = \frac{N-1}{N}\widehat{s}_{xy} = \frac{N-1}{N}\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_{\mathbf{s}})(y_i-\overline{y}_{\mathbf{s}})$$

è uno stimatore corretto della covarianza σ_{xy}. Il termine correttivo $(N-1)/N$, dovuto al fatto che la popolazione di riferimento ha numerosità finita N, è in genere trascurabile a meno che N non sia molto piccolo ($(N-1)/N \approx 1$), per cui si ha $\widehat{\sigma}_{xy} \approx \widehat{s}_{xy}$.

Esempio 3.8. Come esempio numerico consideriamo ancora la popolazione di 1500 famiglie del file `cultura.txt`. Da essa si è selezionato un campione ssr di $n = 100$ unità, i cui dati sono riportati nel file `campione1.txt`. Un modo molto semplice per valutare l'associazione tra spese per attività culturali e reddito disponibile potrebbe consistere nello stimare la covarianza tra questi due caratteri. Sulla base dei dati campionari, è immediato verificare che lo stimatore $\widehat{\sigma}_{xy}$ assume il valore 5972950.52 (si osservi che, nel caso in esame, è $\sigma_{xy} = 6762093.71$). □

3.8 Stima di rapporti

Un problema che di frequente si incontra nella pratica applicativa è quello di stimare un rapporto tra due grandezze. Per capire come sorga questo problema, si consideri la popolazione di 1500 famiglie degli Esempi 3.4–3.7, in cui si osservano il numero di componenti, il reddito netto e le spese per attività culturali di ciascuna famiglia. Accanto alla spesa media familiare, una grandezza di interesse è la spesa media *individuale*, data dal rapporto tra la spesa totale sostenuta dall'intera popolazione e il numero totale di individui della popolazione stessa. In simboli:

$$\textit{spesa media individuale} = \frac{\sum_{i=1}^{1500}\textit{spesa della famiglia } i}{\sum_{i=1}^{1500}\textit{numero di componenti della famiglia } i}\,. \quad (3.39)$$

La costruzione di uno stimatore della (3.39) è un caso speciale di *stima di rapporto*, che può essere formalmente descritto nei termini che seguono. Data una popolazione finita di N unità, supponiamo che siano definiti due caratteri, $\mathcal{X}$, $\mathcal{Y}$. Indichiamo, al solito, con x_i, y_i le modalità assunte rispettivamente da $\mathcal{X}$, $\mathcal{Y}$ in corrispondenza della unità $i$ma della popolazione. Il problema è quello di stimare il rapporto:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i}{\frac{1}{N}\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x} \tag{3.40}$$

in cui μ_x, μ_y sono rispettivamente le medie di $\mathcal{Y}$ e di $\mathcal{X}$ nell'intera popolazione. Come evidenziato dalla (3.40), il quoziente R non è altro che il rapporto tra le medie dei due caratteri $\mathcal{Y}$ e $\mathcal{X}$.

Supponiamo di selezionare un campione $\mathbf{s}$ sulla base di un disegno ssr di ampiezza n. Per ogni unità campionaria, si osservano le modalità x_i, y_i dei due caratteri. Uno stimatore "naturale" di R, suggerito dall'intuizione, è il rapporto tra le medie campionarie dei due caratteri:

$$\widehat{R} = \frac{\overline{y}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}} = \frac{\frac{1}{n}\sum_{i\in\mathbf{s}} y_i}{\frac{1}{n}\sum_{i\in\mathbf{s}} x_i}. \tag{3.41}$$

La distribuzione di probabilità dello stimatore (3.41) è molto difficile da studiare, per una semplice ragione: sia il numeratore che il denominatore di $\widehat{R}$ variano al variare del campione $\mathbf{s}$. Questo fa sì che, in generale, non si possano usare i metodi impiegati per studiare media e varianza della media campionaria. Quello che faremo nel seguito è cercare di ottenere un'espressione approssimata sia per la distorsione di $\widehat{R}$, sia per la sua varianza e per il suo errore quadratico medio.

Proposizione 3.7. *Se il disegno campionario è ssr di numerosità n, valgono le seguenti relazioni:*

$$E[\widehat{R}] \approx R \tag{3.42}$$

$$V(\widehat{R}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{\mu_x^2} \tag{3.43}$$

$$MSE(\widehat{R}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{\mu_x^2}. \tag{3.44}$$

Prima di dimostrare la Proposizione 3.7, è bene discutere brevemente l'ordine di grandezza degli errori di approssimazione insiti nelle (3.42) - (3.44). Come risulterà chiaro dal seguito della presente sezione, gli errori di approssimazione in (3.42) - (3.44) sono tanto più piccoli quanto più grande è la numerosità campionaria n. Più precisamente, per la (3.42) l'errore di approssimazione è dell'ordine di grandezza di $1/n$, mentre per le (3.43), (3.44) è di ordine più piccolo di $1/n$. Questo significa che al crescere di n l'errore di approssimazione in (3.42) decresce alla velocità di $1/n$, mentre gli errori in (3.43), (3.44) decrescono più rapidamente di $1/n$.

Dimostrazione. Anzitutto, si può scrivere

$$\begin{aligned}\widehat{R} - R &= \frac{\overline{y}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}} - R \\ &= \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}} \\ &= \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x} + (\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}) \left(\frac{1}{\overline{x}_{\mathbf{s}}} - \frac{1}{\mu_x} \right). \end{aligned} \tag{3.45}$$

In secondo luogo, con uno sviluppo di Taylor nel punto μ_x si ottiene

$$\frac{1}{\overline{x}_{\mathbf{s}}} = \frac{1}{\mu_x} - \frac{1}{\mu_x^2} (\overline{x}_{\mathbf{s}} - \mu_x) + Resto \tag{3.46}$$

dove il termine *Resto* è di ordine inferiore a $(\overline{x}_{\mathbf{s}} - \mu_x)$. Dalla (3.45) e (3.46) si ha pertanto

$$\widehat{R} = R + \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x} \left(1 - \frac{1}{\mu_x} (\overline{x}_{\mathbf{s}} - \mu_x) + Resto \right). \tag{3.47}$$

Al crescere di n, la varianza di $\overline{x}_{\mathbf{s}}$ tende a 0, per cui la distribuzione di probabilità di $\overline{x}_{\mathbf{s}}$ tende a concentrarsi attorno a μ_x. Equivalentemente, $\overline{x}_{\mathbf{s}} - \mu_x$ tende con alta probabilità a diventare sempre più piccolo, a concentrarsi attorno a 0. Per n "abbastanza grande", quindi, il termine $\overline{x}_{\mathbf{s}} - \mu_x$ tende a diventare piccolo, e "trascurabile" rispetto a 1. Lo stesso vale per *Resto*, che è di ordine inferiore rispetto a $\overline{x}_{\mathbf{s}} - \mu_x$. Questo significa che per n "sufficientemente grande" il termine $(\overline{x}_{\mathbf{s}} - \mu_x)/\mu_x^2 + Resto$ può essere trascurato nella (3.47), e si può scrivere in via approssimata

$$\widehat{R} \approx R + \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x}. \tag{3.48}$$

Dalla (3.48) si ottiene *in primis*

$$\begin{aligned} E[\widehat{R}] &\approx R + \frac{E[\overline{y}_{\mathbf{s}}] - RE[\overline{x}_{\mathbf{s}}]}{\mu_x} \\ &= R + \frac{\mu_y - R\mu_x}{\mu_x} \\ &= R, \end{aligned}$$

il che prova la (3.42). Per quanto concerne le (3.43), (3.44), è sufficiente osservare che, sempre per la (3.48),

$$\begin{aligned} V(\widehat{R}) &\approx V\left(R + \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x}\right) \\ &= \frac{1}{\mu_x^2} V(\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}) \\ &= \frac{1}{\mu_x^2} \{V(\overline{y}_{\mathbf{s}}) + R^2 V(\overline{x}_{\mathbf{s}}) - 2RC(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}})\} \\ &= \frac{1}{\mu_x^2} \left(\frac{1}{n} - \frac{1}{N}\right) (S_y^2 + R^2 S_x^2 - 2RS_{xy}) \end{aligned}$$

e che

$$\begin{aligned} MSE(\widehat{R}) &= V(\widehat{R}) + (E[\widehat{R}] - R)^2 \\ &\approx \frac{1}{\mu_x^2} \left(\frac{1}{n} - \frac{1}{N}\right) (S_y^2 + R^2 S_x^2 - 2RS_{xy}) + (R - R)^2 \\ &= \frac{1}{\mu_x^2} \left(\frac{1}{n} - \frac{1}{N}\right) (S_y^2 + R^2 S_x^2 - 2RS_{xy}). \end{aligned}$$

□

Per valutare, sia pure in modo rozzamente euristico, l'ordine di grandezza di $\overline{x}_{\mathbf{s}} - \mu_x$, calcoliamo il valore atteso di $|\overline{x}_{\mathbf{s}} - \mu_x|$. Si ha:

$$E\left[|\overline{x}_{\mathbf{s}} - \mu_x|\right] \leqslant \sqrt{E\left[(\overline{x}_{\mathbf{s}} - \mu_x)^2\right]} = \sqrt{V(\overline{y}_{\mathbf{s}})} = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} S_x$$

per cui $\overline{x}_{\mathbf{s}} - \mu_x$ è essenzialmente di ordine $1/\sqrt{n}$. Il termine *Resto*, come detto, è di ordine inferiore a $\overline{x}_{\mathbf{s}} - \mu_x$, per cui

$$Resto = \textit{quantità di ordine inferiore a } \frac{1}{\sqrt{n}}.$$

Complessivamente, la (3.47) si può riscrivere come

$$\widehat{R} = R + \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x} \left(1 + \textit{quantità di ordine } \frac{1}{\sqrt{n}}\right).$$

Poiché anche $\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}$ è di ordine $1/\sqrt{n}$ (basta tener conto che il suo valore atteso è nullo, e ripetere parola per parola il ragionamento fatto per $\overline{x}_{\mathbf{s}} - \mu_x$) dalla (3.47) si ottiene

$$\widehat{R} = R + \frac{\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}}}{\mu_x} + \textit{quantità di ordine } \frac{1}{n}. \tag{3.49}$$

Tenendo conto di quanto sopra detto, dalla (3.49) si ottiene

$$E[\widehat{R}] = R + \textit{quantità dell'ordine di } \frac{1}{n}.$$

Per quanto riguarda la varianza di $\widehat{R}$, si può provare (Esercizio 3.12) che

$$V(\widehat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{\mu_x^2} + \textit{quantità di ordine } \frac{1}{\sqrt{n^3}}.$$

Infine, dai risultati finora visti si ottiene (Esercizio 3.13) la seguente relazione per l'errore quadratico medio di $\widehat{R}$:

$$MSE(\widehat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{\mu_x^2} + \textit{quantità di ordine } \frac{1}{\sqrt{n^3}}.$$

La (3.43) si può anche esprimere in forma differente, più semplice. In effetti, da

$$\begin{aligned}
& S_y^2 + R^2 S_x^2 - 2RS_{xy} \\
&= \frac{1}{N-1} \sum_{i=1}^{N} \{(y_i - \mu_y)^2 + R^2 (x_i - \mu_x)^2 - 2R(y_i - \mu_y)(x_i - \mu_x)\} \\
&= \frac{1}{N-1} \sum_{i=1}^{N} \{(y_i - \mu_y) - R(x_i - \mu_x)\}^2 \\
&= \frac{1}{N-1} \sum_{i=1}^{N} (y_i - Rx_i)^2
\end{aligned}$$

si ha subito la seguente espressione, equivalente alla (3.43):

$$V(\widehat{R}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\mu_x^2} \left\{ \frac{1}{N-1} \sum_{i=1}^{N} (y_i - Rx_i)^2 \right\}. \tag{3.50}$$

L'espressione approssimata (3.50) può essere anche utilizzata per costruire uno stimatore di $V(\widehat{R})$. Con gli stessi ragionamenti della Sezione 3.3, e sostituendo gli incogniti R, μ_x rispettivamente con $\widehat{R}$ e $\overline{x}_{\mathbf{s}}$, si può costruire il seguente stimatore di $V(\widehat{R})$:

$$\widehat{V}(\widehat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\overline{x}_{\mathbf{s}}^2} \left\{ \frac{1}{n-1} \sum_{i \in \mathbf{s}} (y_i - \widehat{R}x_i)^2 \right\}. \tag{3.51}$$

Un'espressione alternativa dello stimatore (3.51) è fornita nell'Esercizio 3.14.

Usando infine l'approssimazione normale, si possono anche costruire intervalli di confidenza approssimati per il parametro di interesse R. Usando la solita notazione, infatti, è immediato verificare che

$$\left[\widehat{R} - z_{\alpha/2} \sqrt{\widehat{V}(\widehat{R})},\ \widehat{R} + z_{\alpha/2} \sqrt{\widehat{V}(\widehat{R})} \right]$$

è un intervallo di confidenza per R, al livello (approssimato) $1 - \alpha$.

Esempio 3.9. Si consideri la popolazione di 1500 famiglie del *file* `cultura.txt` (cfr. Esempio 3.5). Da tale popolazione viene selezionato un campione ssr di $n = 30$ famiglie, di cui si osservano reddito annuo (carattere $\mathcal{X}$) e spese annue per attività culturali (carattere $\mathcal{Y}$). Le etichette delle famiglie-campione, assieme ai relativi redditi e spese, sono riportati qui sotto.

Etichetta unità	*Reddito annuo* (x_i)	*Spese attiv. culturali* (y_i)	$y_i - \widehat{R}x_i$
1059	38900.00	8.09	–676.56
1100	43500.00	1244.10	478.5
1120	16800.00	25.80	–269.88
1147	71300.00	2516.20	1261.32
1261	54100.00	365.80	–586.36
1347	26800.00	509.90	38.22
1359	17800.00	178.00	–135.28
1389	32000.00	320.00	–243.2
1393	49000.00	490.00	–372.4
1407	45100.00	467.10	–326.66
1441	64500.00	1612.50	477.3
22	31000.00	310.00	–235.6
240	38000.00	802.20	133.4
274	43100.00	1303.90	545.34
320	37100.00	384.20	–268.76
396	39000.00	780.00	93.6
399	40800.00	816.00	97.92
448	21300.00	316.00	–58.88
599	74900.00	2621.50	1303.26
633	18300.00	189.50	–132.58
643	17600.00	78.40	–231.36
67	35400.00	185.90	–437.14
732	24700.00	521.00	86.28
733	24800.00	131.30	–305.18
743	44200.00	1528.30	750.38
750	53500.00	732.10	–209.5
756	63100.00	941.00	–169.56
821	29100.00	611.80	99.64
866	32700.00	74.10	–501.42
966	38700.00	345.90	–335.22

Le media campionaria del reddito è $\overline{x}_{\mathbf{s}} = 38903.33$ (Euro), mentre la spesa media campionaria per attività culturali è pari a $\overline{y}_{\mathbf{s}} = 686.57$. Questo implica che lo stimatore della frazione di reddito mediamente devoluta in spese per attività culturali è eguale a

$$\widehat{R} = \frac{\overline{y}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}} = \frac{686.57}{38903.33} = 0.0176$$

ovvero è del 1.76%.

Per costruire un intervallo di confidenza per R, iniziamo con l'osservare che

$$\frac{1}{29}\sum_{i\in\mathbf{s}}(y_i - \widehat{R}x_i)^2 = 233855.9$$

da cui si ottiene

$$\sqrt{\widehat{V}(\widehat{R})} = \sqrt{\left(\frac{1}{30} - \frac{1}{1599}\right)\frac{233855.9}{38903.33^2}} = 0.002248\,.$$

Pertanto, tenendo conto che $z_{0.005} = 2.576$, si ottiene per R l'intervallo di confidenza, approssimato al livello 0.99, $[0.0118, 0.02339] = [1.18\%, 2.33\%]$. □

3.9 L'effetto del disegno: aspetti di base*

L'effetto del disegno svolge un ruolo importante in diversi problemi propri del campionamento da popolazioni finite, tra cui, come si vedrà nei capitoli successivi, la scelta del numero di unità del campione. Nella presente sezione ci limiteremo soltanto ad una trattazione di base. Aspetti più avanzati saranno invece considerati nei capitoli successivi.

Supponiamo di voler stimare la media μ_y di una popolazione finita. Una metodologia molto semplice, trattata a fondo nelle precedenti sezioni consiste nell'usare la coppia (*disegno semplice senza ripetizione*, *media campionaria*). Indichiamo con n la numerosità del campione, e con $V(\overline{y}_{\mathbf{s}};\, ssr)$ la varianza della media campionaria. Consideriamo poi un secondo disegno campionario $(\mathcal{S},\, p)$, che seleziona lo stesso numero n di unità elementari, e un secondo stimatore t della media della popolazione. Una qualunque coppia (*disegno campionario*, *stimatore*) costituisce una *strategia campionaria*. Indichiamo con $MSE(t;\, p)$ l'errore quadratico medio dello stimatore t quando usato in coppia con il disegno $(\mathcal{S},\, p)$.

L'*effetto del disegno* (*design effect*: Kish (1965)) è definito come

$$Deff\,(p,\, t) = \frac{MSE(t;\, p)}{V(\overline{y}_{\mathbf{s}};\, ssr)}. \tag{3.52}$$

L'interesse della (3.52) è che permette di valutare la maggiore o minore efficienza della strategia (*disegno* $(\mathcal{S},\, p)$, *stimatore* t) rispetto alla coppia (*disegno ssr*, *media campionaria*). Come già sottolineato, per aver senso il confronto deve essere effettuato a parità di numero di unità elementari che formano il campione. Inoltre, la nozione di effetto del disegno può facilmente estendersi anche alla stima di parametri differenti dalla media campionaria.

L'effetto del disegno (3.52) è una misura della precisione persa o guadagnata dall'utilizzo del disegno $(\mathcal{S},\, p)$ rispetto ad un disegno semplice senza ripetizione. Poiché l'effetto del disegno dipende sia dal disegno di campionamento

che dallo stimatore utilizzato, è evidente che in una stessa rilevazione stime relative a parametri diversi della popolazione (media, totale, proporzione, etc.) possono essere caratterizzate da effetti del disegno diversi.

Se l'effetto del disegno (3.52) è minore di 1 (così che $MSE(t;\, p) < V(\overline{y}_{\mathbf{s}};\, ssr)$) si ha un guadagno di precisione rispetto al campionamento casuale semplice. Viceversa, se la (3.52) è maggiore di 1 ($MSE(t;\, p) < V(\overline{y}_{\mathbf{s}};\, ssr)$) si ha una perdita di precisione. Quanto più la quantità (3.52) è superiore a uno, tanto più la strategia campionaria basata sul disegno semplice senza ripetizione è preferibile alla strategia $((\mathcal{S},\, p),\, t)$.

Un concetto strettamente connesso a quello dell'effetto del disegno è quello di *dimensione campionaria efficace*. Formalmente, se la quantità $Deff\,(p,\, t)$ data dalla (3.52) è nota, ad esempio perché si conosce l'effetto del disegno da una realizzazione precedente dell'indagine o si ricava da un'indagine simile, si ha che la dimensione campionaria (denominata dimensione campionaria efficace) affinché la strategia (*disegno ssr*, *media campionaria*) sia caratterizzata dallo stesso livello di precisione della strategia (*disegno* $(\mathcal{S},\, p)$, *stimatore* t) con numerosità campionaria n. In simboli:

$$n_{eff}(p,\, t) = \frac{n}{Deff\,(p,\, t)}. \tag{3.53}$$

3.10 Il disegno semplice con ripetizione

Il disegno campionario semplice con ripetizione (disegno scr, d'ora in avanti) di dimensione n è definito come segue:

- lo spazio dei campioni è l'insieme di tutte le n-ple ordinate di unità non necessariamente distinte (disposizioni con ripetizione) della popolazione;
- tutti i campioni hanno la stessa probabilità di essere selezionati (equiprobabilità dei campioni).

Due campioni sono considerati distinti se contengono unità diverse o se, pur contenendo le stesse unità, sono caratterizzati da un ordine di selezione diverso.

Lo spazio $\mathcal{S}$ dei campioni (di unità) è l'insieme di tutte le disposizioni con ripetizioni di classe n) del tipo $(i_1,\, i_2,\, \ldots,\, i_n)$, in cui i_1 è la *prima* unità del campione, i_2 è la *seconda* unità del campione, e così via. Inoltre, $i_1,\, i_2,\, \ldots,\, i_n$ possono essere unità qualunque della popolazione, senza altre specificazioni. In modo appena più formale, questo significa che

$$\mathcal{S} = \underbrace{I_N \times I_N \times \cdots \times I_N}_{n\ volte} = I_N^n.$$

Per quanto riguarda le probabilità dei campioni, se $\mathbf{s} = (i_1,\, i_2,\, \ldots,\, i_n)$ si ha

$$p(\mathbf{s}) = \frac{1}{N}\,\frac{1}{N}\,\cdots\,\frac{1}{N} = \frac{1}{N^n}.$$

In termini intuitivi, è come se si effettuassero n "prove"; nella prima prova si seleziona i_1, la prima unità del campione, nella seconda prova si seleziona i_2, la seconda unità del campione, e così via. I risultati delle diverse prove sono indipendenti, hanno identica distribuzione, e sono tali che

$$Pr(i_k = i) = \frac{1}{N} \tag{3.54}$$

per ciascuna unità $i = 1, \ldots, N$ e per ciascuna prova $k = 1, \ldots, n$.

Si vede subito che il disegno scr è *ordinato*, *con ripetizioni*, e *ad ampiezza effettiva non costante*. Infatti, poiché sono ammesse le ripetizioni, si ha sempre $\nu(\mathbf{s}) \leqslant n(\mathbf{s})$. L'uguaglianza si verifica se e solo se nel campione $\mathbf{s}$ non vi sono ripetizioni.

Nel campionamento semplice con ripetizione si ammette la possibilità che una qualunque unità della popolazione possa entrare più di una volta nel campione. Di conseguenza, con tale piano di campionamento è possibile selezionare anche campioni di numerosità n superiore alla numerosità della popolazione oggetto di studio. Nella pratica delle indagini il campionamento è sempre effettuato senza ripetizione. Infatti, se una unità è selezionata due volte nel campione, l'unità stessa sarà osservata una sola volta, e i dati ad essa relativi saranno duplicati in sede di elaborazione. Fissata la dimensione del campione, le ripetizioni comportano chiaramente una perdita di informazione rispetto alla possibilità di disporre di dati relativi a unità differenti. D'altra parte, nel campionamento da popolazioni finite la distinzione tra "con" e "senza" ripetizione diventa irrilevante quando la numerosità della popolazione N è elevata in quanto la probabilità di selezionare due o più volte una stessa unità risulta prossima a zero.

Indichiamo ora con y_{i_k} la modalità dell'unità selezionata nella prova k $(= 1, \ldots, n)$. Come conseguenza della (3.54), anche le variabili aleatorie $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$ sono indipendenti e hanno la stessa distribuzione di probabiltà. Essa è riassunta in Tabella 3.2.

Tabella 3.2 Distribuzione delle variabili aleatorie nel disegno *scr* per la prova $k$ma

i_k	*Probabilità*	y_{i_k}
1	$1/N$	y_1
2	$1/N$	y_2
...	...	...
i	$1/N$	y_i
...	...	...
N	$1/N$	y_N

In particolare, dalla Tabella 3.2 si vede subito che

$$\begin{aligned} E[y_{i_k}] &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \mu_y \quad \textit{per ciascun } k = 1, \ldots, n; \end{aligned} \tag{3.55}$$

$$\begin{aligned} V(y_{i_k}) &= \frac{1}{N}\sum_{i=1}^{N} (y_i - \mu_y)^2 \\ &= \sigma_y^2 \quad \textit{per ciascun } k = 1, \ldots, n. \end{aligned} \tag{3.56}$$

Le (3.55), (3.56) mostrano che la variabile aleatoria y_{i_k} (modalità dell'unità osservata nella prova $k$ma) ha media e varianza uguali a quelle della popolazione; inoltre, $i_1, \ldots, i_n$ sono indipendenti.

Per quanto riguarda la stima della media μ_y della popolazione, uno stimatore "molto naturale" è la media campionaria, definita come

$$\overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} y_i = \frac{1}{n}\sum_{k=1}^{n} y_{i_k}.$$

Come mostrato nella successiva proposizione, se il disegno è di tipo scr la media campionaria è uno stimatore corretto della media della popolazione, e la sua varianza è pari alla varianza della popolazione divisa per la numerosità campionaria.

Proposizione 3.8. *Se il disegno campionario è scr, il valore atteso della media campionaria è pari a:*

$$E(\overline{y}_{\mathbf{s}}) = \mu_y \tag{3.57}$$

e la sua varianza è eguale a

$$V(\overline{y}_{\mathbf{s}}) = \frac{\sigma_y^2}{n}. \tag{3.58}$$

Dimostrazione. Per quanto riguarda la (3.57), è sufficiente tener conto che, per la (3.55), si ha

$$\begin{aligned} E[\overline{y}_{\mathbf{s}}] &= E\left[\frac{1}{n}\sum_{k=1}^{n} y_{i_k}\right] \\ &= \frac{1}{n}\sum_{k=1}^{n} E[y_{i_k}] \\ &= \frac{1}{n}\sum_{k=1}^{n} \mu_y \\ &= \mu_y. \end{aligned}$$

Essendo poi le y_{i_k} indipendenti, dalla (3.56) segue che

$$\begin{aligned} V(\overline{y}_{\mathbf{s}}) &= V\left(\frac{1}{n}\sum_{k=1}^{n} y_{i_k}\right) \\ &= \frac{1}{n^2}\sum_{k=1}^{n} V(y_{i_k}) \\ &= \frac{1}{n^2}\sum_{k=1}^{n} \sigma_y^2 \\ &= \frac{\sigma_y^2}{n}. \end{aligned}$$

□

Usando infine le stesse idee della Proposizione 3.3 è facile vedere che la varianza campionaria corretta

$$\widehat{s}_y^2 = \frac{1}{n-1}\sum_{i\in\mathbf{s}}(y_i - \overline{y}_{\mathbf{s}})^2$$

è uno stimatore corretto della varianza della popolazione.

Proposizione 3.9. *Se il disegno campionario è scr, la varianza campionaria corretta è uno stimatore corretto della varianza della popolazione:*

$$E[\widehat{s}_y^2] = \sigma_y^2. \tag{3.59}$$

Di conseguenza, è immediato verificare che

$$\widehat{V} = \frac{\widehat{s}_y^2}{n} \tag{3.60}$$

è uno stimatore corretto di $V(\overline{y}_{\mathbf{s}})$.

Prima di concludere si deve notare che poiché il valore osservato su una generica unità i del campione è assimilabile alla determinazione di una variabile aleatoria y_{i_k}, alle n unità campionarie sono associate n variabili aleatorie $(Y_{i_1}, \ldots, Y_{i_k}, \ldots, Y_{i_n})$ indipendenti e identicamente distribuite (*i.i.d.*). Questa situazione è tipica del'inferenza statistica "classica" da modello. Si noti come, in effetti, in questa sezione siano stati ricavati risultati fondamentali dell'inferenza statistica classica relativi allo stimatore media campionaria in caso di campionamento con osservazioni *i.i.d.* da popolazioni infinite.

Esercizi

3.1. Si consideri una popolazione finita di $N = 5$ unità, le cui unità possiedono modalità $y_1 = 7$, $y_2 = 4$, $y_3 = 5$, $y_4 = 2$, $y_5 = 8$.

a. Scrivere tutti i campioni ssr di numerosità $n = 3$, con le corrispondenti probabilità.
b. Per ognuno dei campioni al punto *a.*, calcolare la media campionaria $\overline{y}_{\mathbf{s}}$. Verificare quindi che si tratta di uno stimatore corretto della media della popolazione, e calcolarne la varianza.
c. Per ognuno dei campioni al punto *a.* calcolare la varianza campionaria corretta $\widehat{s}_y^2$, e verificare che è uno stimatore corretto di S_y^2.

3.2. Calcolare la probabilità che un campione ssr di numerosità n (da una popolazione di N unità) contenga: *a.* l'unità i; *b.* la coppia $i,\ j$ di unità $(j \neq i)$.

3.3. Da una popolazione di $N = 5876$ famiglie si seleziona un campione ssr di 27 famiglie, di cui si registra il numero di componenti. I dati ottenuti sono qui sotto riportati:

3, 1, 2, 4, 3, 1, 4, 4, 3, 1, 4, 2, 1, 3, 3, 5, 4, 3, 2, 2, 4, 1, 5, 5, 4, 2, 4.

a. Stimare il numero totale di individui della popolazione.
b. Costruire un intervallo di confidenza per il numero di persone nella popolazione, al livello 0.99.

3.4. Selezionare dalla popolazione di 1500 famiglie del file `cultura.txt` un campione ssr di ampiezza $n = 100$. Per i due caratteri "reddito netto disponibile annuo" e "spese annue in attività culturali", calcolare la media campionaria e la varianza campionaria corretta, e costruire un intervallo di confidenza per μ_y al livello 0.99.

3.5. Data una popolazione di $N = 3$ unità, si consideri un disegno ssr di ampiezza $n = 3$, e si definisca lo stimatore di μ_y:

$$\widehat{\mu}(\boldsymbol{y}(\mathbf{s})) = \begin{cases} (y_1 + 2y_2)/3 & \textit{se } \mathbf{s} = \{1, 2\} \\ (2y_1 + y_3)/3 & \textit{se } \mathbf{s} = \{1, 3\} \\ (y_2 + 2y_3)/3 & \textit{se } \mathbf{s} = \{2, 3\} \end{cases} .$$

Provare che si tratta di uno stimatore corretto di μ_y, e calcolare la sua varianza.

3.6. Supposto che la media μ_y della popolazione sia nota, si consideri il seguente stimatore della varianza σ_y^2:

$$\widetilde{\sigma}_c^2 = \frac{1}{n} \sum_{i \in \mathbf{s}} (y_i - \mu_y)(y_i - c)$$

essendo c un numero reale arbitrario.

a. Provare che $\widetilde{\sigma}_c^2$ è uno stimatore corretto di σ_y^2, qualunque sia c reale.
b. Calcolare la varianza di $\widetilde{\sigma}_c^2$.

Suggerimento. Usare la regola di estensione con $t_i(y_i) = (y_i - \mu_y)(y_i - c)$.

c. Verificare che l'efficienza di $\widetilde{\sigma}_c^2$ è massima se $c = (\sum_{i=1}^{N}(y_i - \mu_y)^3/N)/\sigma_y^2 + \mu_y$.

3.7. In un centro di calcolo per studenti vi sono in totale 100 PC. Per controllare lo stato delle tastiere, i due tecnici del centro decidono di procedere in questo modo.

– *Tecnico* 1: osserva tutte le 100 tastiere del centro di calcolo, e registra il fatto che 45 tastiere hanno tutti i tasti funzionanti.
– *Tecnico* 2: osserva un campione ssr di 16 tastiere, e per ognuna di esse registra il numero di tasti mal funzionanti. I dati che egli ottiene sono qui sotto riportati:

$$1,\ 3,\ 0,\ 0,\ 1,\ 1,\ 2,\ 0,\ 1,\ 2,\ 0,\ 1,\ 2,\ 0,\ 3,\ 0.$$

a. Stimare il numero totale di tasti mal funzionanti usando solo i dati del tecnico 2.
b. Stimare il numero totale di tasti mal funzionanti usando sia i dati del tecnico 1 che quelli del tecnico 2.
c. Quale delle due stime ci si aspetta che sia più precisa?

3.8. Da una lista di 2560 studenti iscritti alla Facoltà di Scienze Statistiche si seleziona un campione ssr di numerosità $n = 90$. Di essi, 40 hanno conseguito nella prova di esame di *Tecniche di campionamento* un voto inferiore a 25/30, mentre 50 hanno conseguito un voto almeno pari a 25/30.

a. Stimare la proporzione P_A di studenti che, nella popolazione di riferimento, ha conseguito un voto almeno pari a 25/30.
b. Costruire un intervallo di confidenza per P_A al livello 0.90.

3.9. Due aspiranti sindaci, Romolo e Remo, si contendono la vittoria in un comune di 5000 elettori. Per essere eletti, bisogna guadagnare un numero di voti superiore al 50% degli elettori. Per valutare le proprie possibilità di vittoria, Romolo fa selezionare campione ssr di 300 elettori. 200 degli intervistati dichiarano che voteranno per lui.

a. Costruire un intervallo di confidenza, al livello 0.95, per la proporzione di elettori che voteranno per Romolo.
b. Il portavoce dello *staff* di Romolo afferma che "sulla base dei dati campionari, c'è una netta evidenza che Romolo verrà eletto". Siete d'accordo?

3.10. Si consideri una popolazione finita di $N = 7$ unità, le quali possiedono modalità $y_1, \ldots, y_7$. Da tale popolazione si (i) vuole selezionare un campione di $n = 4$ unità, e (ii) con i dati da esso forniti stimare la media $\mu_y = (y_1 + \cdots + y_7)/7$ della popolazione.

a. Lo statistico Pietro propone che si usi un disegno ssr per selezionare le 4 unità del campione, e la media campionaria per stimare μ_y.

b. Lo statistico Paolo propone che:
- per selezionare le unità campionarie si usi un disegno campionario in cui lo spazio dei campioni è formato da

$$\mathbf{s}_1 = \{1, 2, 4, 7\},\ \mathbf{s}_2 = \{1, 2, 5, 6\},\ \mathbf{s}_3 = \{1, 3, 4, 6\},$$
$$\mathbf{s}_4 = \{1, 3, 5, 7\},\ \mathbf{s}_5 = \{2, 3, 4, 5\},\ \mathbf{s}_6 = \{2, 3, 6, 7\},$$
$$\mathbf{s}_7 = \{4, 5, 6, 7\}$$

  con

$$p(\mathbf{s}_1) = p(\mathbf{s}_2) = \cdots = p(\mathbf{s}_7) = \frac{1}{7};$$

- per stimare la media della popolazione sia impiegata la media campionaria.

Quale delle due proposte è da preferire?

3.11. Si supponga di voler stimare la funzione di ripartizione $F(y)$ (con y fissato) di una popolazione finita di numerosità N (cfr. Sezione 3.6). Dalla popolazione si estrae un campione ssr di numerosità n. Provare che:

a. lo stimatore ottenuto con la regola di estensione (cfr. Sezione 3.6) assume la forma

$$\widehat{F}_n(y) = \frac{\#\ \textit{di unità del campione}\ \mathbf{s}\ \textit{tali che}\ y_i \leqslant y}{n};$$

b. lo stimatore $\widehat{F}_n(y)$ è corretto, e la sua varianza è uguale a $\left(\frac{1}{n} - \frac{1}{N}\right) F(y)\,(1 - F(y))$.

3.12. Con riferimento alla(3.49), provare che $V(\widehat{R}) = V(\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}})/\mu_x^2 +$ *quantità di ordine* $1/\sqrt{n^3}$.

Suggerimento. $V(\widehat{R}) = E[((\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}})/\mu_x +$ *quantità di ordine* $1/n)^2] = V((\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}})/\mu_x) +$ *quantità di ordine* $1/\sqrt{n} \times$ *quantità di ordine* $1/n$.

3.13. Dedurre dall'Esercizio 3.12 che $MSE(\widehat{R}) = V(\overline{y}_{\mathbf{s}} - R\overline{x}_{\mathbf{s}})/\mu_x^2 +$ *quantità di ordine* $1/\sqrt{n^3}$.

3.14. Provare che lo stimatore (3.51) si può anche esprimere come

$$\widehat{V}(\widehat{R}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{\overline{x}_{\mathbf{s}}^2} \{\widehat{s}_y^2 + \widehat{R}^2 \widehat{s}_x^2 - 2\widehat{R}\widehat{s}_{xy}\}.$$

Suggerimento. Usare le (3.12), (3.37) nella (3.43).

3.15. Da una popolazione di $N = 3400$ aziende agricole si seleziona, mediante campionamento ssr, un campione di ampiezza $n = 31$. Per ciascuna azienda-campione si osservano le modalità di due caratteri: la superficie posseduta $\mathcal{X}$ (in are) e quella effettivamente utilizzata $\mathcal{Y}$ (sempre in are). I dati ottenuti sono qui di seguito riportati.

Sup. posseduta	28792	1800	5153	940	102984	13255	4048	700	95	423	2100
Sup. utilizzata	11986	1700	4083	300	77805	10250	4018	600	95	239	2090
Sup. posseduta	850	18585	6697	163770	950	15690	22	1200	8500	619	
Sup. utilizzata	820	18370	6500	30509	880	15622	20	200	8300	447	
Sup. posseduta	52000	20600	5084	18900	5000	948	460	185	20	4230	
Sup. utilizzata	25000	19300	4354	16050	4900	685	340	120	20	4080	

Stimare la proporzione di superficie posseduta che viene mediamente utilizzata, e costruire per tale grandezza un intervallo di confidenza al livello 0.90.

4

Scelta della numerosità campionaria nel campionamento semplice

4.1 Aspetti introduttivi

Fino ad ora è sempre stata assunta come data *a priori* la numerosità campionaria. Ora, in una qualunque rilevazione statistica la numerosità del campione che si utilizza è uno degli elementi più importanti, per diverse ragioni.

1. La numerosità campionaria, come visto nel Capitolo 3, ha una diretta influenza sulla *precisione* della stima della media della popolazione. In effetti, dalla relazione

$$MSE(\overline{y}_{\mathbf{s}}) = V(\overline{y}_{\mathbf{s}}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$$

   si arguisce subito che *quanto più elevata è la numerosità campionaria n, tanto più piccolo è l'errore quadratico medio della media campionaria $\overline{y}_{\mathbf{s}}$*, e quindi *tanto più preciso è $\overline{y}_{\mathbf{s}}$*. Da questo punto di vista, conviene scegliere una numerosità campionaria elevata, in modo da avere una buona precisione di stima.
2. Selezionare unità di una popolazione e (soprattutto) osservare le loro modalità è un'operazione costosa. Poiché le risorse finanziarie disponibili per effettuare una rilevazione statistica sono in genere limitate, *dal punto di vista dei costi conviene che la numerosità campionaria sia piccola.*

I due requisiti 1, 2 sono in contrasto tra loro. Nel prosieguo di questo capitolo adotteremo un approccio che farà (quasi esclusivamente) riferimento al requisito 1 di precisione di stima. Solo nella Sezione 4.2 verrà brevemente delineato un approccio che cerca di inserire i requisiti 1 e 2 in un'unica funzione che combina assieme precisione di stima e costo di rilevazione.

L'idea-guida su cui si fonda gran parte del presente capitolo può essere riassunta in maniera molto semplice. Supponiamo di voler stimare la media μ_y di una popolazione, e di selezionare un campione ssr di numerosità n. Come stimatore di μ_y consideriamo la media campionaria $\overline{y}_{\mathbf{s}}$. Lo stimare μ_y con $\overline{y}_{\mathbf{s}}$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_4,

comporta un *errore di stima*, pari (in valore assoluto) a $|\overline{y}_{\mathbf{s}} - \mu_y|$. Il primo elemento da esplicitare è il *margine di errore* che si considera tollerabile. In dettaglio, si deve fissare un *valore di soglia* $t \geqslant 0$ tale che:

- gli errori di stima al di sotto di t sono considerati "tollerabili";
- gli errori di stima al di sopra di t sono considerati "non tollerabili".

Il valore da dare a t dipende dal tipo di indagine che si conduce, e, in generale, viene fissato dallo statistico.

L'obiettivo ideale sarebbe quello di riuscire a determinare la numerosità campionaria in modo che l'errore di stima non sia mai superiore alla soglia t. Questo, però, è di fatto impossibile, e per una ragione molto semplice. Poiché la media campionaria è una variabile aleatoria, tale sarà anche l'errore di stima $|\overline{y}_{\mathbf{s}} - \mu_y|$. I valori che essa assume dipendono, in generale, dalle modalità y_i, che sono incognite. Tutto quello che si può pretendere è che la variabile aleatoria $|\overline{y}_{\mathbf{s}} - \mu_y|$ sia al di sotto della soglia t *con una data probabilità*, diciamo $1-\alpha$. Anche quest'ultima grandezza viene fissata dallo statistico che pianifica l'indagine campionaria.

Ricapitolando, il requisito 1 è formalizzato richiedendo che l'errore di stima non ecceda una soglia t con una probabilità almeno pari a $1-\alpha$. In simboli:

$$Pr(\textit{errore di stima} \leqslant t) \geqslant 1 - \alpha \tag{4.1}$$

con t e α prefissati in sede di pianificazione della rilevazione. L'idea di base, molto semplice e naturale, è quella di determinare (e usare) la *più piccola* numerosità campionaria che assicura la (4.1). Ovviamente, quest'ultima equivale a

$$Pr(\textit{errore di stima} > t) \leqslant \alpha \tag{4.2}$$

per cui la minima numerosità campionaria per cui vale la (4.1) è uguale alla più piccola numerosità campionaria per cui si ha la (4.2).

In quasi tutto il presente capitolo, come accennato all'inizio, non si farà riferimento al costo di rilevazione in fase di determinazione della numerosità campionaria da utilizzare. Questo non significa però che la considerazione dei costi sia irrilevante. Al contrario, essi giocano un ruolo decisivo. In genere, per effettuare una rilevazione campionaria viene fissato un *budget*, diciamo C, che rappresenta il massimo ammontare spendibile per la rilevazione, e costituisce un limite economico invalicabile. Se la numerosità campionaria determinata in modo che valga la (4.1) (o la (4.2), che è lo stesso) comporta una spesa superiore a C, significa che sono stati fissati dei requisiti di precisione di rilevazione incompatibili con il vincolo economico. In tal caso sarà necessario o fissare dei nuovi requisiti di precisione, meno stretti dei precedenti, oppure dedicare alla rilevazione campionaria maggiori risorse finanziarie.

4.2 Scelta della numerosità campionaria per la stima di proporzioni

Il caso più semplice da affrontare è quello della stima di una proporzione. La situazione è quella già descritta nella Sezione 3.5. Sia P_A la proporzione di unità della popolazione che possiedono l'attributo A, e sia $\widehat{p}_A$ la corrispondente proporzione campionaria di unità che possiedono l'attributo A. Come già visto nella Sezione 3.5, $\widehat{p}_A$ è essenzialmente una media campionaria, e se il disegno usato è ssr di ampiezza n si ha $E[\widehat{p}_A] = P_A$ e $V(\widehat{p}_A) = \{(N-n)/(N-1)\}$ $P_A(1-P_A)/n$.

Conformemente a quanto detto nella precedente sezione, il problema è ora quello di determinare la numerosità campionaria n in maniera tale che l'errore (assoluto) di stima $|\widehat{p}_A - P_A|$ sia superiore ad una soglia t con una probabilità non maggiore di α. In altre parole, bisogna determinare il più piccolo n tale che

$$Pr\left(|\widehat{p}_A - P_A| > t\right) \leqslant \alpha \tag{4.3}$$

con $t > 0$ e α fissati a priori. Per semplificare la trattazione, si può usare l'approssimazione normale per la distribuzione di probabilità di $\widehat{p}_A$. Precisamente, assumeremo che la distribuzione di probabilità di $\widehat{p}_A$ sia approssimata da una normale di media P_A e varianza $\{(N-n)/(N-1)\}P_A(1-P_A)/n$. Questo equivale ad assumere che la v.a. *standardizzata*

$$\frac{\widehat{p}_A - P_A}{\sqrt{\frac{N-n}{N-1}\frac{P_A(1-P_A)}{n}}}$$

abbia, in via approssimata, distribuzione normale standard $N(0,\,1)$. Con questo tipo di approssimazione si ha

$$\begin{aligned} Pr\left(|\widehat{p}_A - P_A| > t\right) &= Pr\left(\left|\frac{\widehat{p}_A - P_A}{\sqrt{\frac{N-n}{N-1}\frac{P_A(1-P_A)}{n}}}\right| > \frac{t}{\sqrt{\frac{N-n}{N-1}\frac{P_A(1-P_A)}{n}}}\right) \\ &= Pr\left(|N(0,\,1)| > \frac{\sqrt{n}t}{\sqrt{P_A(1-P_A)}}\sqrt{\frac{N-1}{N-n}}\right) \\ &= 2Pr\left(N(0,\,1) > \frac{\sqrt{n}t}{\sqrt{P_A(1-P_A)}}\sqrt{\frac{N-1}{N-n}}\right) \end{aligned}$$

a causa della simmetria della distribuzione normale. Con l'approssimazione introdotta, la (4.3) diventa:

$$Pr\left(N(0,\,1) > \frac{\sqrt{n}t}{\sqrt{P_A(1-P_A)}}\sqrt{\frac{N-1}{N-n}}\right) \leqslant \frac{\alpha}{2} \tag{4.4}$$

e il problema è quello di determinare il più piccolo valore di n per cui vale la (4.4). Usando la consueta simbologia, si ha

$$Pr\left(N(0,\,1) > z_{\alpha/2}\right) = \frac{\alpha}{2} \tag{4.5}$$

e dal confronto tra (4.4) e (4.5) si desume che la (4.4) vale se e solo se

$$\frac{\sqrt{n}t}{\sqrt{P_A(1-P_A)}}\sqrt{\frac{N-1}{N-n}} \geqslant z_{\alpha/2}. \tag{4.6}$$

In altri termini, il più piccolo n per cui vale la (4.4) è null'altro che il più piccolo n per cui vale la (4.6). Ma questo significa che deve valere la relazione

$$\frac{\sqrt{n}t}{\sqrt{P_A(1-P_A)}}\sqrt{\frac{N-1}{N-n}} = z_{\alpha/2}$$

dalla quale si ricava

$$\frac{nt^2}{P_A(1-P_A)} = \frac{N-n}{N-1}z^2_{\alpha/2}$$

e quindi

$$\begin{aligned} n &= \frac{\frac{N}{N-1}z^2_{\alpha/2}}{\frac{t^2}{P_A(1-P_A)} + \frac{z^2_{\alpha/2}}{N-1}} \\ &= \frac{\frac{P_A(1-P_A)z^2_{\alpha/2}}{t^2}}{1 + \frac{1}{N}\left(\frac{P_A(1-P_A)z^2_{\alpha/2}}{t^2} - 1\right)}. \end{aligned} \tag{4.7}$$

Il valore di n espresso dalla (4.7) dipende da t, α, $P_A(1-P_A)$, e da N. Ora, è facile verificare che la (4.7) possiede il seguente comportamento (Esercizio 4.1):

- decresce al crescere di t;
- decresce al crescere di α (si osservi che $z_{\alpha/2}$ *decresce* al crescere di α);
- cresce al crescere di $P_A(1-P_A)$;
- cresce al crescere di N (purché sia $P_A(1-P_A)z^2_{\alpha/2}/t^2 > 1$).

Queste affermazioni sono conformi all'intuizione. Infatti, un piccolo valore di t significa ammettere quasi solo errori di stima piccoli, e ciò è possibile solo al prezzo di un'elevata numerosità campionaria. Nello stesso modo, in vista della (4.2), un piccolo valore di α significa essere disposti ad accettare errori di stima elevati solo con bassa probabilità, e questo, ancora, è possibile solo per numerosità campionarie abbastanza elevate. Infine, il termine $P_A(1-P_A)$ è la varianza della popolazione, dalla quale, come visto, dipende la varianza di $\widehat{p}_A$ (si ricordi che $\widehat{p}_A$ è una media campionaria). Quanto più elevata è

$P_A(1-P_A)$, tanto più alta è la varianza di $\widehat{p}_A$, e quindi tanto più è impreciso lo stimatore $\widehat{p}_A$. Affinché $\widehat{p}_A$ raggiunga un dato livello di precisione, sarà necessario utilizzare un'elevata numerosità campionaria.

La (4.7) dipende principalmente da t, da α, e da P_A, mentre è molto meno forte l'influenza della numerosità N della popolazione, almeno quando questa è abbastanza grande. Un'idea di questo fatto è data dalla Tabella 4.1, in cui sono riportate le numerosità campionarie ottenute dalla (4.7) per differenti valori di P_A, α, t, N.

Il termine $P_A(1-P_A)$ è incognito, in quanto è incognito P_A (dopotutto, si tratta proprio del parametro da stimare). Per dare alla formula (4.7) un'utilità pratica, è necessario disporre di una stima preliminare di P_A, diciamo p_{A0}, da

Tabella 4.1 Numerosità campionarie per diversi valori di P_A, α, t, N

				$P_A = 0.1$		
		$\alpha = 0.01$			$\alpha = 0.05$	
N	$t = 0.025$	$t = 0.05$	$t = 0.1$	$t = 0.025$	$t = 0.05$	$t = 0.1$
3000	725	221	59	467	132	34
5000	802	228	59	498	135	34
10000	872	233	59	524	136	34
50000	938	238	59	547	138	34
100000	947	238	59	550	138	34
∞	956	239	60	553	138	35
				$P_A = 0.3$		
		$\alpha = 0.01$			$\alpha = 0.05$	
N	$t = 0.025$	$t = 0.05$	$t = 0.1$	$t = 0.025$	$t = 0.05$	$t = 0.1$
3000	1279	470	133	903	291	79
5000	1542	502	136	1026	303	79
10000	1823	528	137	1143	313	80
50000	2134	551	139	1258	321	81
100000	2181	554	139	1274	322	81
∞	2230	557	139	1291	323	81
				$P_A = 0.5$		
		$\alpha = 0.01$			$\alpha = 0.05$	
N	$t = 0.025$	$t = 0.05$	$t = 0.1$	$t = 0.025$	$t = 0.05$	$t = 0.1$
3000	1409	544	157	1016	341	93
5000	1734	586	161	1176	357	94
10000	2098	622	163	1332	370	95
50000	2521	655	165	1490	381	96
100000	2586	659	166	1513	383	96
∞	2654	664	166	1537	384	96

usare al posto di P_A nella (4.7). La stima preliminare p_{A0} proviene da informazioni extra-campionarie disponibili sulla popolazione oggetto di studio. Ad esempio, potrebbe essere una stima ottenuta con un "piccolo" campione (ssr) appositamente selezionato dalla popolazione (*campione pilota*) o mediante una rilevazione campionaria in un periodo precedente, un valore congetturale fornito da esperti, o altro ancora. Discuteremo più approfonditamente questi aspetti nella Sezione 4.3.

Se non si dispone di una stima preliminare di P_A, una posizione cautelativa (ed anche un po' pessimista) consiste nel sostituire a $P_A(1-P_A)$ il suo valore *massimo*. Tenendo conto della proprietà c, si è in questo modo garantiti che, qualunque sia il valore di P_A, la (4.4) è soddisfatta.

In assenza di informazioni di qualunque tipo, questo significa sostituire a $P_A(1-P_A)$ il massimo valore che esso può assumere, al variare di P_A. È facile provare (Esercizio 4.2) che tale valore massimo è pari a 1/4. Chiaramente la scelta cautelativa comporta una numerosità campionaria più elevata rispetto a quella che si otterrebbe in presenza di informazioni *a priori*.

In questo modo si ottiene il seguente valore per la numerosità campionaria:

$$n_{max} = \frac{\frac{z^2_{\alpha/2}}{4t^2}}{1 + \frac{1}{N}\left(\frac{z^2_{\alpha/2}}{4t^2} - 1\right)}. \tag{4.8}$$

Se la numerosità N della popolazione è abbastanza elevata, il termine

$$\frac{1}{N}\left(\frac{z^2_{\alpha/2}}{4t^2} - 1\right)$$

è praticamente trascurabile, per cui la (4.8) si riduce a

$$n'_{max} = \frac{z^2_{\alpha/2}}{4t^2}. \tag{4.9}$$

In vista della proprietà $d.$, la (4.9) è essenzialmente il massimo valore che può assumere la (4.8), al variare di N.

Talvolta, è noto *a priori* che la proporzione P_A non può superare una dato valore π_{A0}: $P_A \leqslant \pi_{A0}$. Se $\pi_{A0} \leqslant 1/2$, allora si verifica facilmente (Esercizio 4.3) che il valore massimo di $P_A(1-P_A)$ è $\pi_{A0}(1-\pi_{A0})$. Simili considerazioni si possono fare se è noto a priori che $P_A \geqslant \pi_{A0}$. In tutti questi casi, è assai facile modificare la (4.8) per determinare la numerosità campionaria da utilizzare.

Esempio 4.1. Consideriamo ancora la popolazione di 1500 unità del *file* `cultura.txt`, già vista nel Capitolo 3. Si vuole stimare (similmente a quanto visto nel capitolo precedente) la proporzione di individui che spendono ogni anno per cultura più di 1000 Euro. L'obiettivo è che l'errore di stima non superi il 5% con probabilità almeno pari al 98%. Questo significa che deve essere $t = 0.05$ e $1-\alpha = 0.98$, ossia $\alpha = 0.02$.

In assenza di informazioni su P_A, e dato che la numerosità della popolazione è piuttosto contenuta, la numerosità campionaria da usare va determinata in base alla (4.8). Essendo $z_{\alpha/2} = z_{0.01} = 2.326$, si ottiene

$$n = \frac{\frac{2.326^2}{4 \times 0.05^2}}{1 + \frac{1}{1500}\left(\frac{2.326^2}{4 \times 0.05^2} - 1\right)} = 398.$$

Questa numerosità campionaria può essere ridotta non poco se si possiedono informazioni extra-campionarie. Supponiamo che sia noto *a priori* che non più del 30% delle famiglie della popolazione spenda più di 1000 euro per attività culturali. Questo significa che $P_A \leqslant 0.3$, e in tal caso il valore massimo del prodotto $P_A(1-P_A)$ è $0.3(1-0.3) = 0.21$. Utilizzando questo valore nella (4.7), si ottiene

$$n = \frac{\frac{0.21 \times 2.326^2}{0.05^2}}{1 + \frac{1}{1500}\left(\frac{0.21 \times 2.326^2}{0.05^2} - 1\right)} = 349. \qquad \square$$

Fino ad ora si è determinato n in maniera tale che l'*errore assoluto* di stima $|\widehat{p}_A - P_A|$ non superi una data soglia con probabilità (almeno) $1-\alpha$. A volte, però, è di maggior interesse che sia l'*errore relativo* $|\widehat{p}_A - P_A|/P_A$ a soddisfare tale requisito. In altre parole si vuole determinare n in modo tale che

$$Pr\left(\frac{|\widehat{p}_A - P_A|}{P_A} > u\right) \leqslant \alpha.$$

Questa relazione si può riscrivere come

$$Pr\left(|\widehat{p}_A - P_A| > uP_A\right) \leqslant \alpha$$

e il suo confronto con la (4.3) mostra che è esattamente dello stesso tipo, con uP_A al posto di t. Con gli stessi ragionamenti che hanno portato alla (4.7), semplicemente sostituendo t con uP_A, si ottiene per n il valore

$$n = \frac{\left(\frac{1}{P_A} - 1\right)\frac{z^2_{\alpha/2}}{u^2}}{1 + \frac{1}{N}\left\{\left(\frac{1}{P_A} - 1\right)\frac{z^2_{\alpha/2}}{u^2} - 1\right\}}. \tag{4.10}$$

Ovviamente, essendo P_A incognito, per poter utilizzare in pratica la (4.10) occorre disporre di una sua stima preliminare, o comunque di una qualche informazione *a priori* che consenta di delimitarne il valore massimo e/o il minimo. Osserviamo inoltre che, per le stesse ragioni ricordate dianzi, la (4.10) è una funzione crescente di $1/P_A - 1$, ossia è una funzione decrescente di P_A.

Esempio 4.2. Si consideri ancora la popolazione di 1500 unità del *file* `cultura.txt` (cfr. Esempio 4.1). Si vuole stimare la proporzione di individui che spendono ogni anno per cultura più di 1000 Euro, in maniera tale

che che l'errore di stima non superi il 25% di P_A, con probabilità almeno pari al 95%. Si supponga anche di sapere, ad esempio da indagini precedenti, che almeno il 5% delle famiglie della popolazione spendono ogni anno più di 1000 Euro in attività culturali.

Formalmente, bisogna determinare n in modo tale che

$$Pr\left(|\widehat{p}_A - P_A| \leqslant 0.15 P_A\right) \geqslant 0.95$$

con $u = 0.25$ e $\alpha = 0.05$, e noto che $P_A \geqslant 0.05$. Come detto, la (4.10) è una funzione decrescente di P_A, per cui raggiunge il suo massimo per $P_A = 0.05$. Pertanto, il valore di n richiesto, tenendo conto che $z_{0.025} = 1.96$, sarà

$$n = \frac{\left(\frac{1}{0.05} - 1\right)\frac{1.96^2}{0.25^2}}{1 + \frac{1}{1500}\left\{\left(\frac{1}{0.05} - 1\right)\frac{1.96^2}{0.25^2} - 1\right\}} = 657. \qquad \square$$

4.3 Scelta della numerosità campionaria per la stima di medie

Il problema della scelta della numerosità campionaria per stimare la media μ_y di un generico carattere si tratta, in linea di principio, in maniera simile a quanto visto nel caso di proporzioni. Al solito, supponiamo che il disegno usato sia ssr, e che per stimare μ_y si impieghi la media campionaria $\overline{y}_{\mathbf{s}}$. L'errore (assoluto) di stima è pari a $|\overline{y}_{\mathbf{s}} - \mu_y|$. L'obiettivo è sempre quello di determinare la numerosità campionaria n in modo tale che $|\overline{y}_{\mathbf{s}} - \mu_y|$ sia al di sotto di una soglia t con probabilità (almeno) pari a $1 - \alpha$. In simboli:

$$Pr\left(|\overline{y}_{\mathbf{s}} - \mu_y| > t\right) \leqslant \alpha \tag{4.11}$$

con $t > 0$ e α prefissati. Anche adesso, per rendere le cose sufficientemente semplici, si approssima la distribuzione di probabilità di $\overline{y}_{\mathbf{s}}$ con una normale di media μ_y e varianza $(1/n - 1/N)S_y^2$. Ovviamente, ciò equivale a dire che, in via approssimata, la v.a. standardizzata

$$\frac{\overline{y}_{\mathbf{s}} - \mu_y}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}}$$

abbia distribuzione normale standard $N(0, 1)$. Similmente a quanto visto nella sezione precedente si ha allora, posto $f = n/N$ (si ricordi che f è la *frazione sondata*),

$$\begin{aligned} Pr\left(|\overline{y}_{\mathbf{s}} - \mu_y| > t\right) &= Pr\left(\left|\frac{\overline{y}_{\mathbf{s}} - \mu_y}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}}\right| > \frac{t}{\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}}\right) \\ &= Pr\left(|N(0, 1)| > t\frac{\sqrt{n}}{\sqrt{1 - f}S_y}\right) \\ &= 2Pr\left(N(0, 1) > t\frac{\sqrt{n}}{\sqrt{1 - f}S_y}\right). \end{aligned}$$

L'uso dell'approssimazione normale, quindi, porta a riscrivere la (4.11) come:

$$Pr\left(N(0,\,1) > t\frac{\sqrt{n}}{\sqrt{1-f}S_y}\right) \leqslant \frac{\alpha}{2}. \tag{4.12}$$

Il problema da risolvere consiste nell'individuare il più piccolo valore di n per cui vale la (4.12). Gli stessi ragionamenti già fatti nel caso di proporzioni mostrano che deve valere la relazione

$$t\frac{\sqrt{n}}{\sqrt{1-f}S_y} \geqslant z_{\alpha/2} \tag{4.13}$$

per cui il più piccolo n per cui vale la (4.12) coincide con il più piccolo n per cui vale la (4.13). Si può quindi scrivere

$$t\frac{\sqrt{n}}{\sqrt{1-f}S_y} = z_{\alpha/2}$$

da cui, con pochi facili calcoli, si ottiene

$$n = \frac{\frac{z^2_{\alpha/2}}{t^2}S_y^2}{1+\frac{1}{N}\frac{z^2_{\alpha/2}}{t^2}S_y^2}. \tag{4.14}$$

È anche immediato verificare che per N "grande" il termine $(z_{\alpha/2}/t)^2S_y^2/N$ diviene virtualmente trascurabile, per cui la (4.14) si riduce a $n = (z_{\alpha/2}/t)^2S_y^2$.

Per la (4.14) si possono fare commenti quasi identici a quelli della Sezione 4.2. In particolare, è immediato verificare che il valore di n dato dalla (4.14) decresce al crescere di t e di α, mentre cresce al crescere di S_y^2 e di N. Esattamente come nel caso della stima di proporzioni, peraltro, la dipendenza della (4.14) dalla numerosità N della popolazione è abbastanza limitata, mentre molto più accentuata è la sua dipendenza da t, α, S_y^2

Naturalmente, per poter effettivamente usare la (4.14) è necessario conoscere S_y^2, ovvero, in sostanza, la varianza della popolazione. Questo tipo di conoscenza, purtroppo, è in genere molto raro (ed anche poco realistico), per cui è necessario mettere in atto qualche accorgimento per stimare S_y^2, o perlomeno per fornire una sua approssimazione. In questa direzione di lavoro vi sono diverse possibilità.

– Si può stimare S_y^2 con un campione preliminare ssr di numerosità n_p abbastanza piccola. È questa la tecnica del *campione pilota*, usata abbastanza spesso per rilevazioni campionarie su scala medio-grande. Il grande vantaggio che deriva dall'uso di un campione pilota è che esso fornisce informazioni preliminari non solo per stimare S_y^2, ma più in generale utili per mettere a punto la rilevazione vera e propria. Lo svantaggio principale dell'uso del campione pilota sta nel suo costo, in genere tutt'altro che trascurabile, che limita il campo di applicazione del metodo.

- Si possono usare, se disponibili, stime ottenute da rilevazioni precedenti sulla stessa popolazione, o su popolazioni simili. Questa tecnica ha il vantaggio di essere economica e facile da applicare. Tuttavia, va usata con cautela, per due ragioni. In primo luogo, bisogna sempre fare attenzione alla qualità della rilevazione statistica da cui è tratta la stima preliminare di S_y^2. Se la rilevazione è basata su un cattivo disegno, o se ha prodotto dati di bassa qualità (ad esempio con severi errori di misura, o molte mancate risposte), anche la stima di S_y^2 sarà scadente. In secondo luogo, il riferirsi a stime condotte su popolazioni ritenute "simili" a quella di interesse va valutato con cautela, perché potrebbero esservi in realtà differenze anche rilevanti. Questo metodo può anche essere usato per scegliere la numerosità del campione pilota, ad es. scegliendo n_p pari ad una frazione "piccola" (diciamo tra il 5% e il 15%) del valore di n che si otterrebbe usando la (4.14) con S_y^2 stimato da rilevazioni precedenti o da popolazioni "simili" a quella di interesse. Una volta stabilito n_p, si procederebbe come specificato al punto a.
- A volte si dispone di informazioni *a priori* che consentono di costruire almeno una limitazione superiore per S_y^2. Il caso più facile è quello in cui sia noto il campo di variazione del carattere $\mathcal{Y}$ di interesse. Ad esempio, se è noto che $a \leqslant y_i \leqslant b$ per ogni unità i della popolazione, si può dimostrare (Esercizio 4.6) che $\sigma_y^2 \leqslant (b-a)^2/4$, da cui $S_y^2 = N\sigma_y^2/(N-1) \leqslant \{(b-a)^2/4\}/, \{N/(N-1)\}$. A meno che N non sia piccolo, al solito, si ha $N/(N-1) \approx 1$, per cui si può in pratica usare la disuguaglianza $S_y^2 \leqslant (b-a)^2/4$, e approssimare S_y^2 con il suo limite superiore $(b-a)^2/4$. Ovviamente, questo tipo di risultato è utile soltanto se la differenza $b-a$ non è particolarmente grande. Un valore molto grande di $b-a$, infatti, darebbe luogo ad un valore di $(b-a)^2/4$ grande, e quindi porterebbe ad un'approssimazione per eccesso di S_y^2 molto rozza. Questa, a sua volta, fornirebbe un valore di n eccessivamente elevato, e quindi farebbe lievitare i costi di rilevazione. Anche questo metodo può essere usato solo per scegliere la numerosità n_p del campione pilota, esattamente come detto al punto precedente.

Esempio 4.3. Si consideri la popolazione di 1570 studenti del *file* `stature.txt`. Si vuole stimare la statura media della popolazione in modo che l'errore di stima sia non superiore a 1.5 cm. con probabilità almeno 0.92. Usando la notazione dianzi introdotta, questo significa (se le stature sono misurate in cm.) che $t = 1.5$ e $\alpha = 0.08$, da cui $z_{0.04} = 1.751$.

L'uso diretto della formula (4.14) è impossibile, in quanto non si conosce la varianza corretta S_y^2 della popolazione. Una prima idea, semplice da attuare, potrebbe essere quella di basarsi sul campo di variazione delle stature. A meno di casi estremi assai speciali, quasi tutti gli individui della popolazione avranno stature comprese, diciamo, tra 160 e 190 cm.. È quindi ragionevole assumere che $a = 160$ e $b = 190$, da cui $(b-a)^2/4 = 225$. L'uso dell'approssimazione $S_y^2 \approx 225$, però, è del tutto fuori luogo, in quanto il vero valore di S_y^2 è 59.89. Se si determinasse la numerosità campionaria n, tramite la (4.14), ponendo

$S_y^2 = 225$, si otterrebbe un valore

$$n = \frac{\frac{1.751^2}{1.5^2}\,225}{1 + \frac{1}{1570}\frac{1.751^2}{1.5^2}\,225} = 257$$

in effetti piuttosto elevato, di oltre tre volte superiore a quello che si otterrebbe usando il vero valore di S_y^2.

Più utile, in questo caso, è il ricorso ad un campione pilota per stimare S_y^2. Nel caso in esame si è selezionato un campione iniziale di $n = 24$ studenti, di cui si sono misurate le stature (in cm.). I dati ottenuti sono qui sotto riportati.

Matricola	*Statura*	*Matricola*	*Statura*	*Matricola*	*Statura*	*Matricola*	*Statura*
AB2383	174	AB1822	174	AB1088	156	AB1223	158
AB1410	170	AB1112	171	AB1132	163	AB1983	157
AB1482	158	AB1811	170	AB1336	181	AB2069	168
AB2363	180	AB1912	172	AB1833	162	AB1303	182
AB1575	186	AB0970	166	AB2385	176	AB1107	167
AB1926	171	AB1230	171	AB1672	165	AB2288	175

Da tale campione si ottiene una varianza campionaria corretta $\widehat{s}_y^2 = 66.91$. Usando tale valore nella (4.14) si ha una numerosità campionaria finale pari a

$$n = \frac{\frac{1.751^2}{1.5^2}\,66.91}{1 + \frac{1}{1570}\frac{1.751^2}{1.5^2}\,66.91} = 86$$

molto più piccola della precedente.

A questo punto, quel che si fa nella pratica applicativa è di selezionare dalla popolazione, sempre mediante disegno ssr, altre $86 - 24 = 62$ unità (differenti da quelle del campione pilota), e di osservarne le stature. In questo modo, si arriva ad un campione di $n = 86$ unità. Questo modo di procedere è accettabile solo in via approssimata, in quanto non è teoricamente corretto. La ragione di ciò è insita nel fatto che i dati del campione pilota sono usati due volte: per determinare la numerosità del campione finale, e come parte dei dati del campione finale stesso. Questo implica che il numero di unità che fanno parte del campione finale (nel nostro caso 86) dipende dalle modalità di alcune delle unità del campione stesso (quelle del campione pilota). A stretto rigore, il campione finale non potrebbe essere considerato come un campione ssr. Tuttavia, a livello approssimato, si può lavorare *come se* esso fosse un campione ssr di 86 unità della popolazione. □

Similmente a quanto detto nel caso di stima di proporzioni, spesso è di interesse che l'errore relativo $|\overline{y}_{\mathbf{s}} - \mu_y|/\mu_y$ sia più grande di una data soglia u con probabilità non superiore a α. Formalmente, questo significa che bisogna determinare n in modo tale che

$$Pr\left(\left|\frac{\overline{y}_{\mathbf{s}} - \mu_y}{\mu_y}\right| > u\right) \leqslant \alpha.$$

Se si riscrive questa relazione come

$$Pr\left(|\overline{y}_{\mathbf{s}} - \mu_y| > u|\mu_y|\right) \leqslant \alpha$$

si desume subito che essa è esattamente dello stesso tipo della (4.11), con $u|\mu_y|$ al posto di t. Sostituendo t con $u|\mu_y|$ nella (4.14), si ottiene per n il valore

$$n = \frac{\frac{z^2_{\alpha/2}}{u^2}\frac{S^2_y}{\mu^2_y}}{1 + \frac{1}{N}\frac{z^2_{\alpha/2}}{u^2}\frac{S^2_y}{\mu^2_y}}. \tag{4.15}$$

Si noti che il termine S^2_y/μ^2_y è pari a $(N/(N-1))\sigma^2_y/\mu^2_y = (N/(N-1))\,CV(y)^2$, essendo $CV(y) = S_y/|\mu_y|$ il coefficiente di variazione della popolazione. A meno che N non sia piccolo, si ha quindi $S^2_y/\mu^2_y \approx CV(y)^2$, per cui la (4.15) si può riscrivere come

$$n = \frac{\frac{z^2_{\alpha/2}}{u^2}\,CV(y)^2}{1 + \frac{1}{N}\frac{z^2_{\alpha/2}}{u^2}\,CV(y)^2}. \tag{4.16}$$

Essendo il coefficiente di variazione $CV(y)$ in genere incognito, valgono per la (4.16) le stesse considerazioni già fatte per la (4.14). Per maggiori approfondimenti sulla stima preliminare di S^2_y o di $CV(y)$ con un campione pilota si rinvia al volume di Cochran (1977), pp. 78–81.

Esempio 4.4. Consideriamo la popolazione di 1570 studenti del *file* `stature.txt` (cfr. Esempio 4.3). Questa volta si vuole stimare la statura media della popolazione, in modo che l'errore di stima sia non superiore al 1% della statura media della popolazione, con probabilità almeno 0.999. Usando la notazione precedente, deve essere $u = 0.01$ e $\alpha = 0.001$, da cui $z_{0.0005} = 3.291$. Per avere qualche informazione sul coefficiente di variazione della popolazione, facciamo riferimento ai dati del campione pilota dell'Esempio 4.3. Da esso si ottiene $\widehat{s}_y = 8.18$, $\overline{y}_{\mathbf{s}} = 169.71$, e quindi un coefficiente di variazione campionario $\widehat{CV}(y) = 0.048$. Utilizzando questo valore nella (4.16), si ottiene che il campione finale deve avere una numerosità almeno pari a

$$n = \frac{\frac{3.291^2}{0.01^2}\,0.048^2}{1 + \frac{1}{1570}\frac{3.291^2}{0.01^2}\,0.048^2} = 215. \qquad \square$$

4.4 Scelta della numerosità campionaria con approccio decisionale*

Come già accennato nella Sezione 4.1, la numerosità campionaria ha influenza diretta su due elementi, entrambi decisivi per una rilevazione statistica. Da

un lato, la rilevazione dovrebbe produrre stime accurate, ovvero con piccolo errore quadratico medio. Quanto più elevato è n, tanto più precisi sono gli stimatori utilizzati. Da questo punto di vista, accrescere il più possibile la numerosità campionaria è ovviamente conveniente. D'altro canto, un'elevata numerosità campionaria comporta alti costi di rilevazione. Dal punto di vista dei costi, pertanto, conviene usare una numerosità campionaria più piccola possibile.

L'idea di base dell'approccio decisionale consiste nel combinare assieme questi due elementi in un'unica funzione, che esprima in termini numerici l'utilità complessiva che ha, per la rilevazione campionaria, una data numerosità n. L'assunto di base, implicito, è che vi siano due elementi di costo per una rilevazione campionaria, entrambi i quali possono essere perfettamente specificati.

1. Un costo $C_S(n)$ che deriva dal fatto che si osserva non tutta la popolazione, ma solo un suo campione di numerosità n. In altri termini, $C_S(n)$ rappresenta il "costo" dovuto all'imprecisione di stima. Poiché all'aumentare della numerosità campionaria aumenta (in genere) la precisione di stima, si può assumere che $C_S(n)$ decresce al crescere di n.
2. Un costo $C_R(n)$ di osservazione di n unità campionarie. Ovviamente, $C_R(n)$ cresce al crescere di n.

I due elementi $C_S(n)$, $C_R(n)$ vengono poi combinati in un'unica funzione di "costo totale", $C_T(n) = C_S(n) + C_R(n)$. L'idea di base, molto naturale, è di usare la numerosità campionaria n che minimizza il costo totale $C_T(n)$.

La difficoltà maggiore consiste nello specificare in modo esplicito gli elementi che concorrono al costo totale. Per quanto riguarda il costo di osservazione $C_R(n)$ di n unità, è ragionevole assumere, perlomeno nei casi più semplici, che esso sia del tipo:

$$C_R(n) = c_0 + c_1 n \tag{4.17}$$

dove c_0 è un costo fisso e c_1 è il costo che si deve sostenere per osservare un'unità.

Molto più difficile è esplicitare il costo $C_S(n)$ derivante dall'imprecisione di stima. In casi molto semplici si può assumere che sia proporzionale all'errore quadratico medio dello stimatore usato, $\overline{y}_{\mathbf{s}}$. In simboli:

$$C_S(n) = \gamma MSE(\overline{y}_{\mathbf{s}}) = \gamma \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 \tag{4.18}$$

dove γ è una costante di proporzionalità positiva.

Combinando assieme le (4.17) e (4.18) si ottiene la funzione di costo totale:

$$C_T(n) = C_S(n) + C_R(n) = \gamma \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + c_0 + c_1 n. \tag{4.19}$$

È facile verificare (Esercizio 4.9) che il valore di n che rende minima la (4.19) è dato da

$$n = \sqrt{\frac{\gamma}{c_1}} S_y. \tag{4.20}$$

Naturalmente, per calcolare effettivamente il valore di n dato dalla (4.20) è necessario disporre perlomeno di una stima preliminare di S_y.

Come osservazione conclusiva, è da rimarcare che l'approccio decisionale alla determinazione della numerosità campionaria, benché elegante e potenzialmente di grande importanza, è molto di rado usato in pratica. La ragione principale consiste nel fatto che è in generale estremamente difficile specificare, anche solo in maniera approssimata, tutti gli elementi di costo che intervengono in una rilevazione statistica, ed in particolare quelli dovuti all'imprecisione di stima.

Esercizi

4.1. Provare che valgono le asserzioni *a.*–*d.* della Sezione 4.2.

Suggerimento. La funzione $f(x) = x/(1 + a(x - 1))$, $0 < a < 1$, $x \geqslant 0$, cresce al crescere di x. La funzione $g(x) = b/(1 + (b - 1)x)$, $b > 1$, $x \geqslant 0$ cresce al crescere di x.

4.2. Verificare che il massimo valore che può assumere $P_A(1 - P_A)$ è $1/4$.

Suggerimento. La funzione $f(x) = x(1 - x)$ è massima per $x = 1/2$.

4.3. Provare che se $P_A \leqslant \pi_{A0}$, con $\pi_{A0} \leqslant 0.5$, allora il valore massimo di $P_A(1 - P_A)$ è $\pi_{A0}(1 - \pi_{A0})$.

4.4. Per valutare la presenza di errori di stampa in un libro di 752 pagine, si decide di selezionare un campione ssr di pagine, e di rilevare gli errori in essa contenuti. L'interesse è nella stima della frazione P_A di pagine che contengono errori di stampa. Da informazioni sulla precedente edizione del libro, è noto che la frazione di pagine con errori non è superiore al 15%. D'altra parte, il tipo di processo produttivo usato non garantisce che meno del 5% delle pagine contengano errori. Determinare la numerosità campionaria n in modo che l'errore di stima di P_A sia non superiore a 0.05 con probabilità (almeno) pari a 0.95.

4.5. Un politico concorre alle elezioni in un collegio di 50000 elettori. Per valutare le sue *chance* di vittoria, decide di effettuare un sondaggio campionario, mediante campionamento ssr. Sulla base dell'andamento delle passate elezioni, e da informazioni avute dalle sezioni di partito, il politico ritiene ragionevole assumere che avrà almeno il 20% dei voti. Determinare la numerosità campionaria n in modo che l'errore di stima non superi di più del 10% la frazione di elettori che voteranno per il politico, con probabilità almeno pari a 0.9.

4.6. Provare che se $a \leqslant y_i \leqslant b$ per ogni $i = 1, \ldots, N$, allora $\sigma_y^2 \leqslant (b-a)^2/4$.

Suggerimento. Si ha $\sigma_y^2 \leqslant \sum_i (y_i - (a+b)/2)^2/N$, e $|y_i - (a+b)/2|$ può al più essere uguale a $(b-a)/2$.

4.7. Con riferimento alla popolazione di 1500 famiglie del *file* `cultura.txt`, determinare la numerosità che dovrebbe avere un campione ssr se si vuole stimare la spesa media per attività culturali con un errore assoluto non superiore a 100 Euro con probabilità (almeno) pari a 0.9. In assenza di informazioni sulla varianza della popolazione, usare un campione pilota.

4.8. Uno psicologo vuole studiare l'abilità linguistica di bambini di 6 anni. A questo scopo decide di selezionare un campione ssr da una popolazione scolastica di 2534 bambini. Ogni unità del campione è sottoposta ad un test, il cui risultato è un numero compreso tra 1 (abilità minima) e 5 (abilità massima). L'obiettivo è la stima del punteggio medio μ_y della popolazione dei 2534 bambini. Determinare la numerosità campionaria necessaria affinché l'errore di stima non superi il 10% di mu_y, con probabilità pari almeno a 0.9.

4.9. Provare che il valore di n che minimizza la (4.19) è $n = \sqrt{\gamma/c_1}\, S_y$.

5

Stima con il metodo della regressione

5.1 L'uso di caratteri ausiliari: aspetti di base

In linea di principio, come più volte sottolineato, disegno campionario e stimatore(i) usati in una rilevazione campionaria sono scelti dallo statistico che progetta la rilevazione. Tale scelta, ovviamente, è effettuata in modo da assicurare agli stimatori usati un'alta efficienza, ovvero un errore quadratico medio piccolo. La scelta del disegno e degli stimatori dipende non solo (com'è ovvio) dal tipo di problema oggetto di studio, ma anche dalle informazioni *a priori* che si posseggono sulla popolazione di interesse. Supponendo, come sempre si farà nella presente trattazione, che il problema essenziale sia la stima della media della popolazione, ci si è fino ad ora esclusivamente concentrati sulla strategia campionaria (*disegno semplice*, *media campionaria*). Si tratta della più elementare tra le strategie campionarie, utile principalmente quando *non* si è in possesso di informazioni *a priori* sulla popolazione di interesse. Il disegno ssr tratta infatti "alla pari", in maniera *simmetrica*, tutte le unità della popolazione. Inoltre, la media campionaria è il più semplice stimatore della media della popolazione che si possa immaginare. La disponibilità di informazioni *a priori* può intervenire a vari livelli, in quanto le informazioni stesse possono essere usate per modificare o lo stimatore usato, o il disegno campionario, o entrambi.

Naturalmente, è necessario in primo luogo precisare cosa si intende per "informazioni *a priori* sulla popolazione di interesse". Vi sono in effetti molti differenti tipi di informazioni che possono essere in possesso dello statistico che progetta una rilevazione campionaria. Qui si considererà un tipo molto semplice di informazione *a priori*, consistente nella conoscenza delle modalità di un carattere $\mathcal{X}$. Precisamente, si supporrà che siano *note*, per *tutte le unità della popolazione*, le modalità x_i, $i = 1, \ldots, N$, di un *carattere ausiliario* $\mathcal{X}$. Ovviamente, questo implica che sono anche noti tutti i parametri statistici

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_5,

dipendenti solo da $\mathcal{X}$, quali la media

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

la varianza (anche nella versione "corretta")

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2,\ S_x^2 = \frac{N}{N-1}\,\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2$$

o altro ancora.

L'informazione ausiliaria rappresentata dalla conoscenza delle modalità di $\mathcal{X}$ può essere utilizzata sia nella costruzione del disegno campionario, sia nella costruzione dello stimatore permettendo di migliorare, a parità di numerosità campionaria, l'efficienza dei risultati.

Nel presente capitolo sfrutteremo tale informazione esclusivamente a livello di stima. In altre parole, è noto il vettore $\boldsymbol{X}_N$ delle modalità etichettate di $\mathcal{X}$ per tutte le unità della popolazione.

Il disegno che si adotterà sarà ancora quello semplice senza ripetizione. Invece, si introdurranno nuovi tipi di stimatori, dipendenti non solo dalle modalità campionarie di $\mathcal{Y}$ (come accade, ad es., per la media campionaria), ma anche dalle modalità di $\mathcal{X}$. L'idea di base è di sfruttare l'eventuale dipendenza (correlazione) tra $\mathcal{X}$ e $\mathcal{Y}$ per ottenere uno stimatore di μ_y più efficiente della media campionaria $\overline{y}_{\mathbf{s}}$.

Esempi in cui è noto un carattere ausiliario $\mathcal{X}$ sono abbastanza frequenti nella pratica applicativa. Si supponga ad esempio di voler stimare la spesa media sostenuta dalle famiglie che vivono in una regione per l'educazione scolastica dei figli. Le liste anagrafiche forniscono non solo un elenco delle famiglie, ma anche, per ognuna di esse, il numero di componenti e le relazioni di parentela. Un caso che si verifica spesso, poi, è quello in cui $\mathcal{X}$ è lo stesso carattere $\mathcal{Y}$ rilevato sulla popolazione in un periodo precedente.

5.2 Lo stimatore alle differenze

In questa sezione si introduce un primo stimatore che sfrutta la conoscenza del carattere ausiliario $\mathcal{X}$, e le cui proprietà sono molto semplici da studiare: lo stimatore alle differenze. Esso è importante non solo di per sé, ma soprattutto perché fornisce la base logica per l'introduzione dello stimatore per regressione.

Sia $\mathbf{s}$ un campione di numerosità n ottenuto mediante disegno ssr. In corrispondenza di ciascuna unità campionaria i, siano y_i, x_i rispettivamente le modalità etichettate di $\mathcal{Y}$ e di $\mathcal{X}$, e siano

$$\overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} y_i,\ \ \overline{x}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in\mathbf{s}} x_i$$

le corrispondenti medie campionarie.

Se c è una costante reale arbitraria, lo *stimatore alle differenze* $\widehat{\mu}_{d,\,c}$ è definito come

$$\widehat{\mu}_{d,\,c} = \overline{y}_{\mathbf{s}} - c(\overline{x}_{\mathbf{s}} - \mu_x).$$

Si noti che l'applicazione di tale stimatore non richiede che siano realmente noti i valori x_i per tutte le unità della popolazione, ma unicamente il valore della media μ_x, e le modalità x_i per le sole unità campionarie.

Proposizione 5.1. *Se il disegno campionario è ssr, lo stimatore alle differenze* $\widehat{\mu}_{d,\,c}$ *è uno stimatore corretto della media della popolazione:*

$$E\left[\widehat{\mu}_{d,\,c}\right] = \mu_y. \tag{5.1}$$

La varianza di $\widehat{\mu}_{d,\,c}$ *è pari a*

$$V(\widehat{\mu}_{d,\,c}) = \left(\frac{1}{n} - \frac{1}{N}\right)(S_y^2 + c^2 S_x^2 - 2cS_{xy}). \tag{5.2}$$

Dimostrazione. Per provare la (5.1) basta osservare che, essendo il disegno ssr, si ha

$$\begin{aligned} E\left[\widehat{\mu}_{d,\,c}\right] &= E\left[\overline{y}_{\mathbf{s}} - c(\overline{x}_{\mathbf{s}} - \mu_x)\right] \\ &= E\left[\overline{y}_{\mathbf{s}}\right] - c\left(E\left[\overline{x}_{\mathbf{s}}\right] - \mu_x\right) \\ &= \mu_y - c(\mu_x - \mu_x) \\ &= \mu_y. \end{aligned}$$

Per quanto riguarda la varianza di $\widehat{\mu}_{d,\,c}$, usando le Proposizioni 3.2, 3.5 si può scrivere

$$\begin{aligned} V(\widehat{\mu}_{d,\,c}) &= V\left(\overline{y}_{\mathbf{s}} - c(\overline{x}_{\mathbf{s}} - \mu_x)\right) \\ &= V\left(\overline{y}_{\mathbf{s}} - c\overline{x}_{\mathbf{s}}\right) \\ &= V(\overline{y}_{\mathbf{s}}) + c^2 V(\overline{x}_{\mathbf{s}}) - 2cC(\overline{x}_{\mathbf{s}}, \overline{y}_{\mathbf{s}}) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + c^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_x^2 - 2c \left(\frac{1}{n} - \frac{1}{N}\right) S_{xy} \end{aligned}$$

da cui segue subito la (5.2). □

Come detto, la costante c è arbitraria. Ovviamente, conviene scegliere il valore di c che rende massima l'efficienza dello stimatore $\widehat{\mu}_{d,\,c}$, ossia il valore di c che minimizza la sua varianza. Derivando la (5.2) rispetto a c, si ha

$$\frac{dV(\widehat{\mu}_{d,\,c})}{dc} = \left(\frac{1}{n} - \frac{1}{N}\right)(2cS_x^2 - 2S_{xy})$$

da cui

$$\frac{dV(\widehat{\mu}_{d,\,c})}{dc} = 0 \text{ } \textit{se e solo se} \text{ } cS_x^2 - S_{xy} = 0.$$

Tenendo anche conto che la derivata seconda di $V(\widehat{\mu}_{d,\,c})$ è positiva, il valore di c che rende minima la (5.2) è pari a:

$$c = \frac{S_{xy}}{S_x^2} = \frac{\sigma_{xy}}{\sigma_x^2} = b_{y/x}$$

ovvero al coefficiente di regressione (nella popolazione) di $\mathcal{Y}$ rispetto a $\mathcal{X}$. Con la posizione $c = b_{y/x}$, lo stimatore alle differenze assume la forma

$$\widehat{\mu}_{d,\,b_{y/x}} = \overline{y}_{\mathbf{s}} - b_{y/x}(\overline{x}_{\mathbf{s}} - \mu_x). \tag{5.3}$$

La sua varianza, in questo caso, è eguale a

$$\begin{aligned} V(\widehat{\mu}_{d,\,b_{y/x}}) &= \left(\frac{1}{n} - \frac{1}{N}\right)\left\{S_y^2 + \frac{S_{xy}^2}{S_x^4}S_x^2 - 2\frac{S_{xy}}{S_x^2}S_{xy}\right\} \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2}\right) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 (1 - \rho_{xy}^2) \end{aligned}$$

dove

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{S_{xy}}{S_x S_y}$$

è il coefficiente di correlazione lineare tra $\mathcal{X}$ e $\mathcal{Y}$. Essendo $-1 \leqslant \rho_{xy} \leqslant 1$, si ha $0 \leqslant \rho_{xy}^2 \leqslant 1$, da cui:

$$V(\widehat{\mu}_{d,\,b_{y/x}}) \leqslant \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 = V(\overline{y}_{\mathbf{s}})$$

e l'uguaglianza tra le due varianze ha luogo se solo se $\rho_{xy} = 0$, ossia se e solo se i due caratteri sono incorrelati. In ogni altro caso, lo stimatore alle differenze "ottimale" (5.3) è *più efficiente* della media campionaria. La differenza tra $V(\overline{y}_{\mathbf{s}})$ e $V(\widehat{\mu}_{d,\,b_{y/x}})$ è tanto più grande quanto più elevato è, in valore assoluto, ρ_{xy}, ossia quanto più forte è il legame lineare tra i due caratteri $\mathcal{X}$ e $\mathcal{Y}$. Se la correlazione è perfetta, $\rho_{xy} = \pm 1$, la varianza dello stimatore alle differenze è nulla.

Per capire in maniera un po' più approfondita la struttura dello stimatore (5.3), consideriamo la retta di regressione, nella popolazione, di $\mathcal{Y}$ rispetto a $\mathcal{X}$: $y = a_{y/x} + b_{y/x}\,x$, con $b_{y/x} = S_{xy}/S_x^2$ e $a_{y/x} = \mu_y - b_{y/x}\mu_x$. Tra le modalità y_i, x_i dei due caratteri $\mathcal{Y}$, $\mathcal{X}$ sussiste la relazione:

$$y_i = a_{y/x} + b_{y/x}\,x_i + e_i, \quad i = 1, \ldots, N$$

in cui i termini $e_i = y_i - a_{y/x} - b_{y/x}\,x_i$, $i = 1, \ldots, N$ sono gli *errori*. Per come il coefficiente $a_{y/x}$ è definito, si vede subito che le y_i si possono scrivere come:

$$y_i = \mu_y + b_{y/x}(x_i - \mu_x) + e_i, \quad i = 1, \ldots, N$$

da cui discende che

$$e_i = (y_i - \mu_y) - b_{y/x}(x_i - \mu_x), \quad i = 1, \ldots, N. \tag{5.4}$$

Le media degli errori nella popolazione è zero:

$$\mu_e = \frac{1}{N}\sum_{i=1}^{N} e_i = 0.$$

Inoltre, è anche facile verificare (Esercizio 5.2) che la covarianza (corretta) tra gli errori e le $\mathcal{X}$ è nulla:

$$S_{xe} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)e_i = 0 \tag{5.5}$$

e che, come conseguenza, la varianza dell'errore è pari a

$$\begin{aligned} S_e^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(e_i - \mu_e)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N} e_i^2 \\ &= S_y^2 - b_{y/x}^2 S_x^2 \\ &= S_y^2(1 - \rho_{xy}^2). \end{aligned} \tag{5.6}$$

La relazione (5.6) mostra che la varianza dello stimatore alle differenze "ottimale" (5.3) si può scrivere come:

$$V(\widehat{\mu}_{d,\, b_{y/x}}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_e^2$$

e quindi che essa dipende essenzialmente dai termini di errore della regressione lineare di $\mathcal{Y}$ rispetto a $\mathcal{X}$. Quanto più piccoli sono gli errori e_i (in valore assoluto), tanto più piccola è la loro varianza S_e^2, tanto più efficiente è $\widehat{\mu}_{d,\, b_{y/x}}$. La massima efficienza si raggiunge quando gli errori e_i sono tutti nulli. In tal caso si ha infatti $S_e^2 = 0$, per cui $\widehat{\mu}_{d,\, b_{y/x}}$ è identicamente uguale alla media μ_y da stimare. È quasi superfluo sottolineare che S_e^2 è tanto più piccola quanto più grande è ρ_{xy}^2, e che si ha $S_e^2 = 0$ se e solo se il coefficiente di correlazione ρ_{xy} è uguale a 1 o a -1, per cui si ritrova per questa via quanto detto in precedenza.

5.3 Lo stimatore per regressione

Lo stimatore alle differenze "ottimale" (5.3), pur avendo eccellenti proprietà, presenta un sostanziale difetto che ne limita enormemente la portata applicativa: richiede che il coefficiente di regressione $b_{y/x}$ sia noto. Ora, in quasi tutti i casi di interesse applicativo tale quantità è incognita, per cui $\widehat{\mu}_{d,\, b_{y/x}}$ non può essere utilizzato. L'idea di base per rimediare a questo inconveniente

è semplice: *stimare $b_{y/x}$ su base campionaria*. Il più semplice stimatore di $b_{y/x}$ è il *coefficiente di regressione campionario* di $\mathcal{Y}$ rispetto a $\mathcal{X}$:

$$\widehat{b}_{y/x} = \frac{\widehat{s}_{xy}}{\widehat{s}_x^2} = \frac{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})\,(y_i - \overline{y}_\mathbf{s})}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})^2}. \tag{5.7}$$

Sostituendo $\widehat{b}_{y/x}$ a $b_{y/x}$ nella (5.3), si ha lo *stimatore per regressione* di μ_y:

$$\widehat{\mu}_{reg} = \overline{y}_\mathbf{s} - \widehat{b}_{y/x}(\overline{x}_\mathbf{s} - \mu_x). \tag{5.8}$$

Dello stimatore per regressione si può anche dare un'interpretazione differente, ma equivalente. Sia $y = a_{y/x} + b_{y/x}x$ la retta di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ nella popolazione. Come ben noto, si ha $b_{y/x} = S_{xy}/S_x^2$ e $a_{y/x} = \mu_y - b_{y/x}\mu_x$, per cui vale la relazione $\mu_y = a_{y/x} + b_{y/x}\mu_x$. Se si conoscessero esattamente i coefficienti di regressione $a_{y/x}$ e $b_{y/x}$, a partire da μ_x si potrebbe determinare esattamente la media μ_y. Il problema è che $a_{y/x}$ e $b_{y/x}$ non sono noti. L'idea, molto naturale, è quella di stimarli su base campionaria. Tramite gli n dati campionari $\{(x_i, y_i);\ i \in \mathbf{s}\}$, si può costruire la retta di regressione campionaria di $\mathcal{Y}$ rispetto a $\mathcal{X}$: $y = \widehat{a}_{y/x} + \widehat{b}_{y/x}x$, con $\widehat{a}_{y/x} = \overline{y}_\mathbf{s} - \widehat{b}_{y/x}\overline{x}_\mathbf{s}$ e $\widehat{b}_{y/x}$ definito nella (5.7). Per $x = \mu_x$ si ottiene lo stimatore per regressione:

$$\widehat{a}_{y/x} + \widehat{b}_{y/x}\,\mu_x = \overline{y}_\mathbf{s} - \widehat{b}_{y/x}(\overline{x}_\mathbf{s} - \mu_x) = \widehat{\mu}_{reg}.$$

Pertanto, *$\widehat{\mu}_{reg}$ è null'altro che l'ordinata della retta di regressione campionaria (di $\mathcal{Y}$ rispetto a $\mathcal{X}$) corrispondente all'ascissa μ_x* (vds. Fig. 5.1).

Il calcolo esatto del valore atteso e della varianza dello stimatore per regressione (5.8) è estremamente difficile, in quanto sia il numeratore che il denominatore di $\widehat{b}_{y/x}$ variano al variare del campione $\mathbf{s}$. Pertanto, a meno di casi molto speciali, il valore atteso di $\widehat{b}_{y/x}$ non è uguale a $b_{y/x}$, così come il valore atteso di $\widehat{b}_{y/x}(\overline{x}_\mathbf{s} - \mu_x)$ non è uguale a zero. Nella prossima sezione studieremo, in modo approssimato e con metodi simili a quelli usati per la stima del rapporto tra due medie, il valore atteso e la varianza dello stimatore per regressione.

Esempio 5.1. Si supponga di voler stimare la produzione di grano di una certa regione, e di non disporre di un elenco delle aziende agricole presenti nella regione stessa. Questa situazione è piuttosto comune in molti paesi in via di sviluppo, in cui non vi sono anagrafi di imprese. Una soluzione semplice ed economica potrebbe essere quella di ripartire la regione in parcelle di territorio delle stesse dimensioni, e di riprendere foto aeree delle stesse. Tali foto aeree dovrebbero permettere di determinare la proporzione di ogni parcella coltivata a grano, e quindi fornirebbero anche una valutazione della produzione di grano di ogni parcella. In questo modo si potrebbe calcolare la produzione media di grano delle parcelle, la quale, moltiplicata per il numero di parcelle in

cui è suddivisa la regione, permetterebbe di valutare la produzione totale di grano della regione. Questo approccio (usato, in diverse varianti, nella pratica applicativa) ha il pregio dell'estrema economicità di rilevazione. L'elemento più problematico consiste nel fatto che il calcolo della produzione delle diverse parcelle ottenuto mediante foto aeree è rozzo e soggetto ad errori. Per questa ragione è opportuno selezionare in una seconda fase un campione di parcelle, ognuna delle quali viene esaminata da un esperto in grado di fornire una valutazione della corrispondente produzione di grano. L'osservazione diretta di esperti fornisce quasi sempre valutazioni molto accurate della produzione di grano delle diverse parcelle, ma d'altra parte è molto costosa, per cui il relativo campione ha numerosità esigua.

Lo stimatore per regressione permette di combinare sia i dati ottenuti da foto aeree, sia quelli provenienti da osservazione diretta a terra. Per capire meglio questo punto, facciamo riferimento ad un esempio numerico. Il *file* `grano.txt` contiene, per un complesso di $N = 2500$ parcelle di terreno, le valutazioni della produzione di grano (in quintali) ottenute da foto aeree e da osservazione a terra, che indichiamo rispettivamente con x_i e y_i, $i = 1, \ldots, 2500$. I valori x_i sono noti, così come la media μ_x. Dei valori y_i si osserva invece un campione ssr di ampiezza $n = 20$. I dati ottenuti sono qui sotto riportati.

Etichetta i	*Ril. aerea* x_i *(q.li)*	*Ril. a terra* y_i *(q.li)*
1744	16	26
1823	26	34
1351	39	33
2031	57	48
1846	53	44
920	68	46
51	44	42
106	71	53
545	12	10
844	61	45
2188	22	27
798	36	32
529	45	37
562	46	40
440	41	31
2380	120	100
2370	32	30
1967	44	45
2432	0	8
1289	49	41

Le medie campionarie di $\mathcal{Y}$ e $\mathcal{X}$ sono pari a $\overline{y}_{\mathbf{s}} = 38.60$ e $\overline{x}_{\mathbf{s}} = 44.10$.

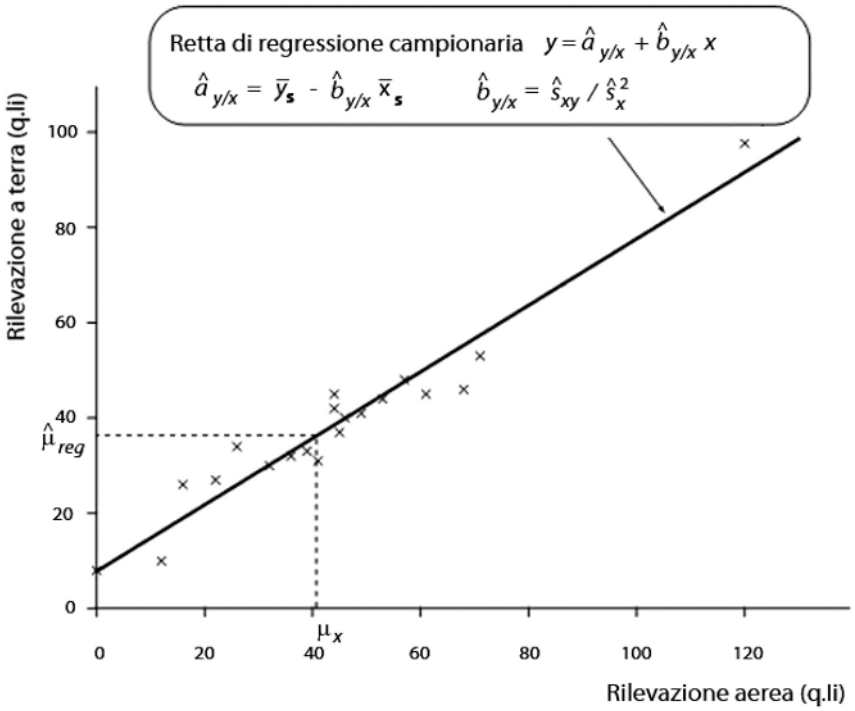


Fig. 5.1 Stimatore per regressione per un campione di $n = 20$ parcelle

Inoltre, da

$$\widehat{s}_x^2 = \frac{1}{19} \sum_{i \in \mathbf{s}} (x_i - \overline{x}_{\mathbf{s}})^2 = 654.94,$$

$$\widehat{s}_{xy} = \frac{1}{19} \sum_{i \in \mathbf{s}} (x_i - \overline{x}_{\mathbf{s}}) (y_i - \overline{y}_{\mathbf{s}}) = 457.09$$

si deduce che $\widehat{b}_{y/x} = 457.09/654.94 = 0.698$. Pertanto, lo stimatore per regressione è pari a

$$\widehat{\mu}_{reg} = 38.60 - 0.698\,(44.10 - 41.12) = 35.51.$$

In Fig. 5.1 è rappresentato graficamente, per l'esempio in esame, il modo in cui lo stimatore per regressione di μ_y è costruito a partire dai dati campionari. □

Un punto molto importante riguarda le circostanze in cui è realmente opportuno usare lo stimatore per regressione (5.8) anziché la media campionaria. In maniera tutto sommato ovvia, ed anche un po' vaga, si può dire $\widehat{\mu}_{reg}$ va

usato al posto di $\overline{y}_{\mathbf{s}}$ nei casi in cui ci si attende da esso una maggiore efficienza di stima. La costruzione di $\widehat{\mu}_{reg}$ vista sopra consente di dare un senso lievemente più preciso a questa affermazione. L'idea di fondo su cui si basa lo stimatore per regressione è di sfruttare la (eventuale) relazione lineare che, a livello di popolazione, intercorre tra i caratteri di interesse $\mathcal{Y}$ e ausiliario $\mathcal{X}$. Quanto più forte è questa relazione, tanto migliore sarà il comportamento di $\widehat{\mu}_{reg}$ in termini di efficienza. La relazione lineare tra $\mathcal{Y}$ e $\mathcal{X}$ è misurata dal coefficiente di correlazione lineare ρ_{xy}, per cui si può concludere che l'uso dello stimatore $\widehat{\mu}_{reg}$ è opportuno se si può ragionevolmente assumere che il coefficiente di correlazione lineare tra $\mathcal{Y}$ e $\mathcal{X}$ sia, in valore assoluto, abbastanza alto. Quello che conta per assicurare allo stimatore per regressione (5.8) delle buone caratteristiche di efficienza è che vi sia una forte relazione *lineare* tra $\mathcal{Y}$ e $\mathcal{X}$. Quando si ha ragione di ritenere che $\mathcal{Y}$ e $\mathcal{X}$ abbiano un debole legame lineare l'uso dello stimatore (5.8) è fuori luogo, in quanto la sua efficienza potrebbe essere più bassa di quella della media campionaria $\overline{y}_{\mathbf{s}}$.

Una seconda considerazione piuttosto importante riguarda la numerosità campionaria. A meno che il coefficiente di correlazione lineare tra $\mathcal{Y}$ e $\mathcal{X}$ non sia molto elevato in valore assoluto, lo stimatore per regressione non fornisce risultati apprezzabili quando la numerosità campionaria è molto piccola. Uno studio empirico su questo punto è in un lavoro di Rao (1969), in cui si analizza, con metodi di simulazione di tipo Monte Carlo, il comportamento dello stimatore per regressione in otto popolazioni naturali, per numerosità campionarie "piccole", dell'ordine di $n = 12$ o meno.

La discussione appena svolta si basa tutta sulla formalizzazione delle relazioni di dipendenza che sussistono tra $\mathcal{Y}$ e $\mathcal{X}$. Un modo naturale, quasi ovvio, di formalizzare e studiare la struttura delle relazioni di dipendenza tra $\mathcal{Y}$ e $\mathcal{X}$ consiste nell'usare un *modello di superpopolazione* , in si cui assume che le y_i non siano semplici numeri, ma piuttosto realizzazioni di variabili aleatorie Y_i, legate alle x_i da un modello di regressione lineare $Y_i = a + bx_i + U_i$, essendo U_i la variabile aleatoria "errore di regressione". Si rientra in questo modo nell'ambito "classico" dei modelli di regressione lineare. L'approccio basato su modelli di superpopolazione è ampiamente usato nel campionamento da popolazioni finite. Dato il livello elementare della presente trattazione, per il momento ci accontentiamo solo di questi brevi cenni.

5.4 Distorsione e varianza approssimate dello stimatore per regressione

L'obiettivo di questa sezione è di studiare, in modo approssimato, il valore atteso e la varianza dello stimatore per regressione (5.8). La tecnica usata per studiare questo problema è molto simile a quella usata nella Sezione 3.8 per studiare in modo approssimato valore atteso e varianza del rapporto di due medie campionarie. Il risultato principale che si otterrà è riassunto nella seguente proposizione.

Proposizione 5.2. *Se il disegno campionario è ssr di numerosità n, valgono le seguenti relazioni:*

$$E[\widehat{\mu}_{reg}] \approx \mu_y \tag{5.9}$$

$$V(\widehat{\mu}_{reg}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1-\rho_{xy}^2) \tag{5.10}$$

$$MSE(\widehat{\mu}_{reg}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1-\rho_{xy}^2). \tag{5.11}$$

Gli errori di approssimazione presenti nelle (5.9) - (5.11) sono tanto più piccoli quanto più grande è la numerosità campionaria n. Come verificheremo più avanti, per la (5.9) l'errore di approssimazione è dell'ordine di grandezza di $1/n$, mentre per le (5.10), (5.11) è di ordine più piccolo di $1/n$. Al crescere di n, pertanto, l'errore di approssimazione in (5.9) decresce alla velocità di $1/n$, mentre gli errori in (5.10), (5.11) decrescono più rapidamente di $1/n$. Questo significa che la Proposizione 5.2 fornisce indicazioni effettivamente utili solo se la numerosità campionaria n è "grande". Per valori piccoli di n, le grandezze $E[\widehat{\mu}_{reg}]$, $V(\widehat{\mu}_{reg})$, $MSE(\widehat{\mu}_{reg})$ possono divergere anche in misura considerevole dai loro valori approssimati che figurano nelle (5.9) - (5.11).

Dimostrazione. Per studiare, sia pure in modo approssimato, le proprietà dello stimatore per regressione, iniziamo con l'osservare che vale la relazione

$$\begin{aligned}
\widehat{b}_{y/x} &= \frac{\frac{1}{n-1}\sum_{i\in\mathbf{s}} y_i(x_i-\overline{x}_\mathbf{s})}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2} \\
&= \frac{\frac{1}{n-1}\sum_{i\in\mathbf{s}}\{\mu_y + b_{y/x}(x_i-\mu_x)+e_i\}(x_i-\overline{x}_\mathbf{s})}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2} \\
&= \frac{\frac{1}{n-1}\left\{\mu_y\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s}) + b_{y/x}\sum_{i\in\mathbf{s}}(x_i-\mu_x)(x_i-\overline{x}_\mathbf{s}) + \sum_{i\in\mathbf{s}} e_i(x_i-\overline{x}_\mathbf{s})\right\}}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2} \\
&= \frac{\frac{1}{n-1}\left\{b_{y/x}\sum_{i\in\mathbf{s}} x_i(x_i-\overline{x}_\mathbf{s}) + \sum_{i\in\mathbf{s}} e_i(x_i-\overline{x}_\mathbf{s})\right\}}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2} \\
&= b_{y/x} + \frac{\frac{1}{n-1}\sum_{i\in\mathbf{s}} e_i(x_i-\overline{x}_\mathbf{s})}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2} \\
&= b_{y/x} + \frac{\widehat{s}_{xe}}{\widehat{s}_x^2}
\end{aligned} \tag{5.12}$$

in cui, come conseguenza dei risultati delle Sezioni 3.3 e 3.7,

$$\widehat{s}_x^2 = \frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})^2, \quad \widehat{s}_{xe} = \frac{1}{n-1}\sum_{i\in\mathbf{s}} e_i(x_i-\overline{x}_\mathbf{s})$$

sono stimatori corretti rispettivamente di S_x^2 e di $S_{xe} = 0$ (5.6).

Si può dimostrare (si veda, ad esempio, Herzel (1982)) che $E[(\widehat{s}_x^2 - S_x^2)^2] = V(\widehat{s}_x^2)$ è dell'ordine di grandezza di $1/n$. Ragionando in modo euristico, non rigoroso, questo implica che $(\widehat{s}_x^2 - S_x^2)^2$ è esso stesso dell'ordine di $1/n$, e quindi che $(\widehat{s}_x^2 - S_x^2)$ è dell'ordine di grandezza di $1/\sqrt{n}$. Nello stesso modo, si vede che anche $(\widehat{s}_{xe} - S_{xe}) = \widehat{s}_{xe}$ è dell'ordine di grandezza di $1/\sqrt{n}$. Dalla (5.12) si conclude quindi che vale la seguente relazione:

$$\begin{aligned} \widehat{b}_{y/x} - b_{y/x} &= \frac{\textit{quantità di ordine } \frac{1}{\sqrt{n}}}{S_x^2 + \textit{quantità di ordine } \frac{1}{\sqrt{n}}} \\ &= \textit{quantità di ordine } \frac{1}{\sqrt{n}}. \end{aligned} \tag{5.13}$$

A sua volta, tenendo anche conto che il termine $(\overline{x}_{\mathbf{s}} - \mu_x)$ è dell'ordine di grandezza di $1/\sqrt{n}$, la (5.13) permette di concludere che lo stimatore per regressione si può esprimere come

$$\begin{aligned} \widehat{\mu}_{reg} &= \overline{y}_{\mathbf{s}} - \left(b_{y/x} + \textit{quantità di ordine } \frac{1}{\sqrt{n}} \right) (\overline{x}_{\mathbf{s}} - \mu_x) \\ &= \overline{y}_{\mathbf{s}} - b_{y/x}(\overline{x}_{\mathbf{s}} - \mu_x) + (\overline{x}_{\mathbf{s}} - \mu_x) \times \textit{quantità di ordine } \frac{1}{\sqrt{n}} \\ &= \overline{y}_{\mathbf{s}} - b_{y/x}(\overline{x}_{\mathbf{s}} - \mu_x) + \textit{quantità di ordine } \frac{1}{n}. \end{aligned} \tag{5.14}$$

Dalla (5.14) è immediato ricavare la (5.9):

$$\begin{aligned} E[\widehat{\mu}_{reg}] &= E\left[\overline{y}_{\mathbf{s}} - b_{y/x}(\overline{x}_{\mathbf{s}} - \mu_x)\right] + E\left[\textit{quantità di ordine } \frac{1}{n}\right] \\ &= \mu_y + \textit{quantità di ordine } \frac{1}{n}. \end{aligned}$$

Usando infine la (5.14) e considerazioni simili a quelle sopra viste, è facile verificare (Esercizi 5.3, 5.4) che

$$V(\widehat{\mu}_{reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1 - \rho_{xy}^2) + \textit{quantità di ordine più piccolo di } \frac{1}{n};$$

$$MSE(\widehat{\mu}_{reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1 - \rho_{xy}^2) + \textit{quantità di ordine più piccolo di } \frac{1}{n}.$$

□

5.5 Stima della varianza dello stimatore per regressione

Come sempre accade quando si stima un parametro (nel nostro caso, la media della popolazione) è necessario anche fornire una stima dell'errore quadratico medio dello stimatore utilizzato. Le relazioni (5.10), (5.11) ci dicono che

la varianza e l'errore quadratico medio di $\widehat{\mu}_{reg}$ si approssimano esattamente nello stesso modo, per cui è sufficiente stimare la varianza di $\widehat{\mu}_{reg}$ per avere uno stimatore del suo errore quadratico medio. Dall'espressione approssimata (5.10) si ha la relazione

$$\begin{aligned} V(\widehat{\mu}_{reg}) &\approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 (1 - S_{xy}^2/(S_x^2 S_y^2)) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \left\{S_y^2 - b_{y/x}^2 S_x^2\right\}. \end{aligned}$$

Sostituendo a S_y^2, S_x^2, $b_{y/x}$ i corrispondenti stimatori $\widehat{s}_y^2$, $\widehat{s}_x^2$, $\widehat{b}_{y/x}$, si ha il seguente stimatore di $V(\widehat{\mu}_{reg})$:

$$\widehat{V}(\widehat{\mu}_{reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\widehat{s}_y^2 - \widehat{b}_{y/x}^2 \widehat{s}_x^2\right\}. \tag{5.15}$$

Un'espressione alternativa dello stimatore (5.15) è fornita nell'Esercizio 5.5.

Se si usa l'approssimazione normale per la distribuzione di probabilità di $\widehat{\mu}_{reg}$, si possono anche costruire intervalli di confidenza approssimati per la media della popolazione. Con la consueta notazione, infatti, è pressoché immediato verificare che

$$\left[\widehat{\mu}_{reg} - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_{reg})},\ \widehat{\mu}_{reg} + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_{reg})}\right]$$

è un intervallo di confidenza per la media μ_y della popolazione, al livello (approssimato) $1 - \alpha$.

Esempio 5.2. Consideriamo ancora l'Esempio 5.1. Come già visto, è $\widehat{s}_x^2 = 654.94$, $\widehat{s}_{xy} = 457.09$, $b_{y/x} = 0.698$. Inoltre, è facile verificare che $\widehat{s}_y^2 = 343.62$, per cui come stima di $V(\widehat{\mu}_{reg})$ avremo la seguente:

$$\widehat{V}(\widehat{\mu}_{reg}) = \left(\frac{1}{20} - \frac{1}{2500}\right) \left\{343.62 - 0.698^2 654.94\right\} = 1.217.$$

Per un livello di confidenza $1-a = 0.96$ si ha dalle tavole della distribuzione normale standard $z_{0.02} = 2.054$, da cui segue che l'intervallo

$$\left[35.51 - 2.054\sqrt{1.217},\ 35.51 + 2.054\sqrt{1.217}\right] = [33.24,\ 37.78]$$

è un intervallo di confidenza approssimato al livello 0.96 per la media μ_y della popolazione. □

Esercizi

5.1. Verificare che lo stimatore $\widehat{\mu}_{d,\,c}$ è più efficiente della media campionaria purché sia $0 < c < b_{y/x}$ se $b_{y/x} > 0$, e $b_{y/x} < c < 0$ se $b_{y/x} < 0$.

5.2. Provare la relazione (5.5).

Suggerimento. Dalla (5.4) discende che $\sum_{i=1}^{N}(x_i - \mu_x)e_i = \sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) - b_{y/x}\sum_{i=1}^{N}(x_i - \mu_x)^2$.

5.3. Verificare che $V(\widehat{\mu}_{reg}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1 - \rho_{xy}^2)$ + *quantità di ordine più piccolo di* $\frac{1}{n}$.

Suggerimento. Usare la (5.14), e tenere conto che $V($*quantità di ordine* $\frac{1}{n})$ e $C(\overline{y}_{\mathbf{s}} - b_{y/x}(\overline{x}_{\mathbf{s}} - \mu_x)$, *quantità di ordine* $\frac{1}{n})$ sono di ordine più piccolo di $1/n$.

5.4. Verificare che $MSE(\widehat{\mu}_{reg}) = V(\widehat{\mu}_{reg})$ + *quantità di ordine più piccolo di* $\frac{1}{n}$.

5.5. Verificare che vale la relazione:

$$\widehat{V}(\widehat{\mu}_{reg}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i\in\mathbf{s}}\{(y_i - \overline{y}_{\mathbf{s}}) - \widehat{b}_{y/x}(x_i - \overline{x}_{\mathbf{s}})\}^2.$$

5.6. Un agricoltore vuole valutare la produzione media di mele di un frutteto, in cui vi sono in totale $N = 150$ meli. Per ogni albero si può dare sia una valutazione "ad occhio" della quantità di mele prodotte (x_i), sia una valutazione precisa (y_i) ottenuta cogliendo le mele dell'albero e pesandole. In base alle valutazioni ad occhio su tutti gli alberi del frutteto, si ha che ogni albero dovrebbe produrre una quantità media di mele pari a 46 kg. Viene poi selezionato un campione ssr di $n = 18$ meli, di ognuno dei quali si pesa la produzione di mele. I risultati sono riportati nella tabella qui di seguito.

Peso ad occhio x_i *(kg)*	*Peso effettivo* y_i *(kg)*	*Peso ad occhio* x_i *(kg)*	*Peso effettivo* y_i *(kg)*
41	41	39	36
37	39	51	54
42	46	44	50
46	40	55	54
48	51	42	40
45	43	50	47
40	35	43	41
52	54	53	53
46	48	49	50

a. Stimare μ_y mediante lo stimatore per regressione.
b. Basandosi sul risultato in (i), costruire un intervallo di confidenza per μ_y al livello 0.95.

6

Stima con il metodo del quoziente

6.1 Aspetti di base: definizione dello stimatore per quoziente

Lo stimatore per quoziente è, come struttura logica, simile a quello per regressione. Poniamoci esattamente nelle condizioni del Capitolo 5, ovvero supponiamo che siano note le modalità $x_1, \ldots, x_N$ che un carattere ausiliario $\mathcal{X}$ assume in corrispondenza delle unità della popolazione. Indichiamo con $\mu_x = \sum_{i=1}^{N} x_i/N$ la media del carattere $\mathcal{X}$ nella popolazione. L'obiettivo, come al solito, è quello di stimare la media μ_y del carattere $\mathcal{Y}$. In tutto il presente capitolo si assumerà che il disegno di campionamento è quello semplice senza ripetizione.

Rispetto a quanto assunto nel Capitolo 5, supponiamo di disporre di un'informazione aggiuntiva: *la retta di regressione di Y rispetto a X passa per l'origine*. Tenendo conto che tale retta deve anche passare per il punto (μ_x, μ_y), se $\mu_x \neq 0$ (come sempre implicitamente assumeremo d'ora in avanti) questo significa che essa ha equazione del tipo:

$$y = \frac{\mu_y}{\mu_x}\, x = R\, x \tag{6.1}$$

in cui $R = \mu_y/\mu_x$ è il rapporto tra la media di $\mathcal{Y}$ e quella di $\mathcal{X}$. Dalla (6.1) si ricava subito l'ovvia relazione

$$\mu_y = R\, \mu_x. \tag{6.2}$$

Se R fosse noto, si potrebbe calcolare esattamente μ_y usando la (6.2). Essendo R incognito, l'idea di base è quella di usare la relazione (6.2) a livello campionario, stimando R tramite il rapporto tra la media campionaria di $\mathcal{X}$ e quella di $\mathcal{Y}$:

$$\widehat{R} = \frac{\overline{y}_s}{\overline{x}_s}.$$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_6,

Si ottiene in questo modo lo *stimatore per quoziente* di μ_y, definito come:

$$\widehat{\mu}_q = \widehat{R}\,\mu_x = \frac{\overline{y}_s}{\overline{x}_s}\,\mu_x. \tag{6.3}$$

Per quanto riguarda l'uso dello stimatore per quoziente, una domanda molto naturale, pressoché scontata, è la seguente: "Quando è opportuno usare lo stimatore per quoziente? La risposta a tale quesito è in sostanza insita nella costruzione dello stimatore stesso.

1. La prima condizione che deve essere verificata è che *la retta di regressione di Y rispetto a $\mathcal{X}$*, almeno in via approssimata, *passi per l'origine*. Quanto più si è vicini a condizioni di questo tipo, tanto più alta è l'efficienza dello stimatore per quoziente.
2. In secondo luogo, vale una considerazione simile a quella fatta per lo stimatore per regressione: lo stimatore per quoziente fornisce risultati tanto migliori quanto più il carattere oggetto di interesse $\mathcal{Y}$ e il carattere ausiliario $\mathcal{X}$ sono correlati.

Queste due condizioni possono essere riassunte dicendo che, in termini di efficienza, lo stimatore per quoziente fornisce buoni risultati quando tra i due caratteri $\mathcal{Y}$ e $\mathcal{X}$ intercorre una relazione di approssimata *proporzionalità*.

Al limite, se tra le modalità di $\mathcal{Y}$ e quelle di $\mathcal{X}$ vi fosse una relazione di esatta proporzionalità, si avrebbe $y_i = R\,x_i$ per ogni unità i della popolazione, e quindi $\overline{y}_s = R\,\overline{x}_s$. A sua volta, questo implicherebbe che $\widehat{R} = R$, e quindi $\widehat{\mu}_q = \mu_y$. In altri termini, se $\mathcal{Y}$ e $\mathcal{X}$ sono esattamente proporzionali, lo stimatore per quoziente diviene identicamente uguale alla media μ_y da stimare.

Queste considerazioni portano in modo semplice ad una importante conclusione: *quanto più si è vicini ad una situazione di proporzionalità tra le y_i e le x_i, tanto migliore è la precisione di stima dello stimatore per quoziente*. È questo il criterio-guida che porta a scegliere di usare lo stimatore per quoziente per stimare μ_y. Se, in base alle informazioni *a priori* di cui si dispone, si può ragionevolmente assumere di essere abbastanza vicini ad una situazione di quasi-proporzionalità tra le modalità dei caratteri $\mathcal{Y}$ e $\mathcal{X}$, l'uso dello stimatore per quoziente (6.3) è opportuno. Se però questo tipo di assunzione non può essere sostenuta, è preferibile non usare lo stimatore per quoziente, in quanto potrebbe dar luogo a severi errori di stima.

A meno che non si sia del tutto sicuri della validità delle assunzioni a priori su cui si basa la scelta dello stimatore per quoziente, è spesso necessario valutare, sia pure rozzamente, la validità di tali assunzioni sulla base dei dati campionari. Un modo molto semplice di operare consiste nello studiare i residui. Posto $\widehat{y}_i = \widehat{R}\,x_i$, definiamo i *residui campionari* rispetto alla retta di equazione $\widehat{y} = \widehat{R}\,x$ come:

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{R}\,x_i, \quad i \in \mathbf{s}. \tag{6.4}$$

Su un diagramma cartesiano rappresentiamo poi le coppie $(x_i, \widehat{e}_i)$, per tutte le n unità del campione $\mathbf{s}$. Se sono valide le assunzioni 1, 2, è intuitivo attendersi

che la media campionaria dei residui sia approssimativamente pari a zero: $\widehat{\overline{e}}_{\mathbf{s}} = \sum_{i \in \mathbf{s}} \widehat{e}_i / n \approx 0$. Il primo controllo da effettuare riguarda pertanto il valore di $\widehat{\overline{e}}_{\mathbf{s}}$, che dovrebbe risultare abbastanza prossimo a zero.

Un'analisi molto più fine della precedente riguarda la retta di regressione campionaria (dei minimi quadrati) dei residui $\widehat{e}_i$ rispetto alle x_i. Sotto le assunzioni 1, 2, essa dovrebbe approssimativamente coincidere con l'asse delle ascisse. Si osservi che quando ciò accade si ha automaticamente $\widehat{\overline{e}}_{\mathbf{s}} \approx 0$. Il secondo tipo di controllo da effettuare riguarda quindi la retta di regressione campionaria delle $\widehat{e}_i$ rispetto alle x_i, che dovrebbe risultare prossima all'asse delle ascisse.

Se le analisi sopra illustrate validano le assunzioni 1, 2, si può usare lo stimatore per quoziente (6.3) con ragionevoli aspettative di buona efficienza.

Esempio 6.1. Un botanico vuole valutare l'età di una foresta di $N = 426$ alberi. Per determinare l'età di un albero vi è un metodo preciso, consistente nel contare il numero di anelli concentrici del fusto (ogni anello corrisponde ad un anno). Questo metodo, però, ha il fondamentale difetto di richiedere il taglio della pianta. Un metodo di valutazione dell'età di un albero più semplice e incruento, ma molto meno accurato, consiste nel misurare il diametro del fusto di un albero. L'intuizione suggerisce che il diametro di un fusto dovrebbe essere grosso modo proporzionale al numero di anelli del fusto stesso. Per valutare l'età media degli alberi della foresta si potrebbe quindi adottare la procedura qui sotto descritta.

- Si misurano i diametri di tutti gli alberi della foresta. In questo modo resta definito, sulla popolazione di alberi, un carattere ausiliario noto.
- Si seleziona un campione ssr di alberi, di cui si contano gli anelli.

Come già accennato, è ragionevole ammettere che vi sia una relazione di approssimata proporzionalità tra età e diametro degli alberi. Risulta quindi sensato utilizzare lo stimatore per quoziente per stimare l'età media μ_y degli alberi della foresta.

Il *file* `alberi.txt` contiene, per gli $N = 426$ alberi della foresta, sia il diametro (in cm.) che l'età (in anni). Indichiamo rispettivamente con x_i e y_i, $i = 1, \ldots, 426$, queste due quantità. I valori x_i sono noti per tutti gli alberi, così come la loro media $\mu_x = 27.23$. Dei valori y_i si osserva invece un campione ssr di ampiezza $n = 24$. I dati ottenuti sono qui sotto riportati.

Etichetta i	*Diametro* x_i *(cm.)*	*Età* y_i *(anni)*	$\widehat{y}_i = \widehat{R}\, x_i$	*Residui* $\widehat{e}_i = y_i - \widehat{y}_i$
270	24.1	76	95.195	−19.195
288	23.9	98	94.405	3.595
344	40.6	164	160.37	3.63
302	25.7	109	101.515	7.485
133	32.5	125	128.375	−3.375
268	25.4	113	100.33	12.67
350	31.2	114	123.24	−9.24
318	19.1	72	75.445	−3.445
352	20.6	85	81.37	3.63
256	22.6	100	89.27	10.73
27	34.5	142	136.275	5.725
90	22.9	94	90.455	3.545
378	31.2	122	123.24	−1.24
124	30.5	125	120.475	4.525
85	31.0	118	122.45	−4.45
246	24.1	90	95.195	−5.195
381	19.6	73	77.42	−4.42
180	32.5	116	128.375	−12.375
296	29.2	124	115.34	8.66
306	26.2	98	103.49	−5.49
385	26.7	113	105.465	7.535
146	31.0	127	122.45	4.55
172	25.4	91	100.33	−9.33
380	31.5	123	124.425	−1.425

Le medie campionarie di $\mathcal{Y}$ e $\mathcal{X}$ sono rispettivamente $\overline{y}_{\mathbf{s}} = 108.83$ e $\overline{x}_{\mathbf{s}} = 27.58$, così che si ha

$$\widehat{R} = \frac{108.83}{27.58} = 3.95.$$

I dati campionari possono essere usati per validare l'assunzione di approssimata proporzionalità tra $\mathcal{Y}$ e $\mathcal{X}$, come illustrato in precedenza. La media campionaria dei residui è pari a $\overline{e}_{\mathbf{s}} = -0.121$, e quindi molto prossima a zero. Per quanto riguarda la retta di regressione dei residui rispetto ai diametri, essa è rappresentata in Fig. 6.1; il coefficiente angolare è pari a 0.022, e l'intercetta a −0.73.

Chiaramente, tale retta è quasi coincidente con l'asse delle ascisse. Tutto ciò permette di affermare che, in via approssimata, vi è una relazione di quasi-proporzionalità tra diametri ed età degli alberi, e giustifica l'uso dello stimatore per quoziente, che risulta pari a

$$\widehat{\mu}_q = \widehat{R}\, \mu_x = 3.95 \times 27.23 = 107.56.$$

In Fig. 6.2 è rappresentato graficamente il modo in cui lo stimatore per quoziente è costruito a partire dalla retta di equazione $\widehat{y} = \widehat{R}\, x$.

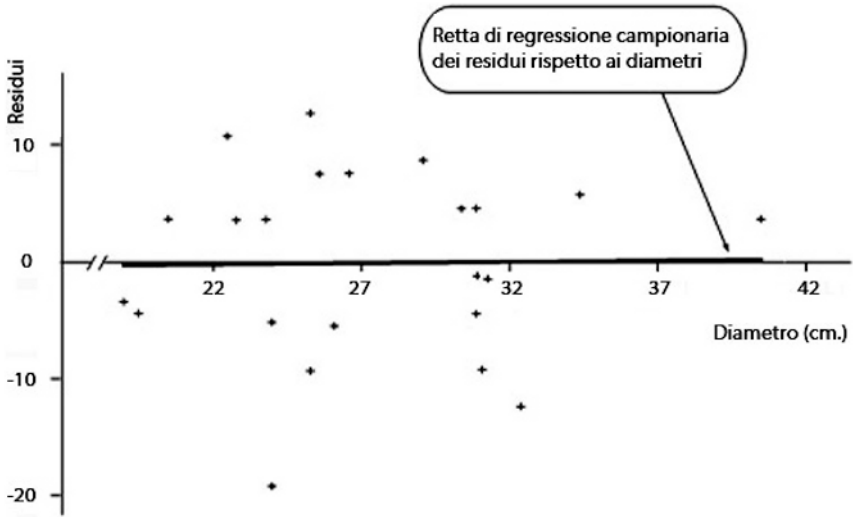


Fig. 6.1 Retta di regressione dei residui rispetto ai diametri degli alberi

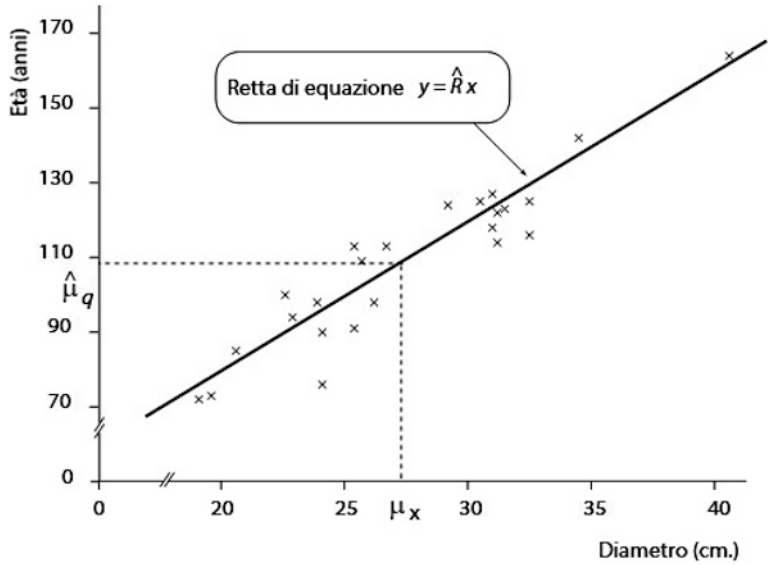


Fig. 6.2 Stimatore per quoziente per un campione di $n = 24$ alberi

È interessante osservare che, a fronte di un valore della media della popolazione pari a $\mu_y = 107.5$, la media campionaria $\overline{y}_{\mathbf{s}}$ è pari a 108.83, e quindi *sovrastima* μ_y. Tramite la relazione di quasi-proporzionalità tra età e diametri degli alberi, lo stimatore per quoziente, che assume il valore $\widehat{\mu}_q = 107.56$, corregge in larga parte questa sovrastima. □

6.2 Distorsione e varianza approssimate dello stimatore per quoziente

Se il disegno usato per selezionare le unità campionarie è di tipo ssr, lo stimatore per quoziente (6.3) è *distorto*. In effetti, come già visto nel Capitolo 3 a proposito della stima di un rapporto di due medie, si ha

$$E[\widehat{R}] = E\left[\frac{\overline{y}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}}\right] \neq \frac{E[\overline{y}_{\mathbf{s}}]}{E[\overline{x}_{\mathbf{s}}]} = \frac{\mu_y}{\mu_x} = R$$

e quindi

$$E[\widehat{\mu}_q] = E[\widehat{R}\,\mu_x] = E[\widehat{R}]\,\mu_x \neq R\mu_x = \mu_y.$$

Della distorsione $B(\widehat{\mu}_q)$ $(= E[\widehat{\mu}_q] - \mu_y)$ dello stimatore per quoziente si può anche fornire la seguente espressione esatta (vds. Esercizio 6.1 per la dimostrazione):

$$B(\widehat{\mu}_q) = -C(\widehat{R}, \overline{x}_{\mathbf{s}}). \tag{6.5}$$

L'espressione (6.5), benché molto elegante, è in pratica di scarsa utilità, in quanto impossibile da calcolare esplicitamente. Risulta quindi molto più utile, perlomeno nel caso del campionamento ssr, cercare di fornire un'espressione approssimata della distorsione e dell'errore quadratico medio dello stimatore per quoziente, sulla falsariga di quanto già visto per lo stimatore per regressione.

Essendo $\widehat{\mu}_q = \widehat{R}\,\mu_x$, le proprietà dello stimatore per quoziente dipendono sostanzialmente da quelle di $\widehat{R} = \overline{y}_{\mathbf{s}}/\overline{x}_{\mathbf{s}}$. In via approssimata, media e varianza di $\widehat{R}$ sono già state studiate nella Sezione 3.8. Si possono quindi sfruttare i risultati già ottenuti. In particolare, dalla Proposizione 3.7 è immediato far discendere i seguenti risultati.

Proposizione 6.1. *Se il disegno campionario è ssr di numerosità n, valgono le seguenti relazioni:*

$$E[\widehat{\mu}_q] = E[\widehat{R}]\,\mu_x \approx R\,\mu_x = \mu_y \tag{6.6}$$

$$V(\widehat{\mu}_q) = V(\widehat{R})\,\mu_x^2 \approx \left(\frac{1}{n} - \frac{1}{N}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{xy}\} \tag{6.7}$$

$$MSE(\widehat{\mu}_q) = V(\widehat{R})\,\mu_x^2 \approx \left(\frac{1}{n} - \frac{1}{N}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{xy}\}. \tag{6.8}$$

L'ordine di grandezza degli errori delle approssimazioni (6.6) – (6.8) dipende dall'ordine di grandezza degli errori che si commettono approssimando $E[\widehat{R}]$, $V(\widehat{R})$, $MSE(\widehat{R})$. Questo è già stato studiato nella Sezione 3.8. Da questa discende subito che:

$$E[\widehat{\mu}_q] = \mu_y + \textit{quantità di ordine } \frac{1}{n}$$

$$V(\widehat{\mu}_q) = \left(\frac{1}{n} - \frac{1}{N}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{xy}\} + \textit{quantità di ordine più piccolo di } \frac{1}{n}$$

$$MSE(\widehat{\mu}_q) = \left(\frac{1}{n} - \frac{1}{N}\right)\{S_y^2 + R^2 S_x^2 - 2RS_{xy}\} + \textit{quantità di ordine più piccolo di } \frac{1}{n}.$$

La formula (6.8) permette di valutare, sia pure in modo approssimato, il guadagno (o la perdita) di efficienza che l'uso dello stimatore per quoziente comporta rispetto alla media campionaria, fermo restando che il disegno campionario è ssr. Si ha infatti:

$$\begin{aligned} MSE(\overline{y}_{\mathbf{s}}) - MSE(\widehat{\mu}_q) &\approx \left(\frac{1}{n} - \frac{1}{N}\right)\left\{S_y^2 - (S_y^2 + R^2 S_x^2 - 2RS_{xy})\right\} \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)(2RS_{xy} - R^2 S_x^2) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) RS_x \left(2\frac{S_{xy}}{S_x S_y} S_y - RS_x\right) \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) RS_x (2\rho_{xy} S_y - RS_x) \end{aligned} \tag{6.9}$$

essendo ρ_{xy} il coefficiente di correlazione lineare tra $\mathcal{X}$ e $\mathcal{Y}$. Dalla (6.9) risulta chiaro che lo stimatore per quoziente è più efficiente della media campionaria se e solo se $2\rho_{xy} S_y - RS_x \geqslant 0$, ovvero se e solo se

$$\rho_{xy} \geqslant \frac{RS_x}{2\,S_y} = \frac{S_x/\mu_x}{2S_y/\mu_y} = \frac{CV(x)}{2\,CV(y)}$$

dove $CV(x)$ e $CV(y)$ sono rispettivamente il coefficiente di variazione di $\mathcal{X}$ e quello di $\mathcal{Y}$.

6.3 Stima della varianza dello stimatore per quoziente

Dalla formula della varianza approssimata (in effetti sarebbe più corretto parlare di errore quadratico medio approssimato) dello stimatore per quoziente si può anche dare un'espressione alternativa. Per rendere più semplice il risultato finale conviene partire da quanto ottenuto nella Sezione 3.8, ed in particolare dalla relazione

$$S_y^2 + R^2 S_x^2 - 2RS_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

dalla quale discende che

$$V(\widehat{\mu}_q) \approx \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{N-1} \sum_{i=1}^{N} (y_i - Rx_i)^2. \tag{6.10}$$

La (6.10) suggerisce un semplice stimatore di $V(\widehat{\mu}_q)$, ottenuto sostituendo l'incognito rapporto $R = \mu_y/\mu_x$ con la sua "controparte campionaria" $\widehat{R} = \overline{y}_{\mathbf{s}}/\overline{x}_{\mathbf{s}}$. Si ha in questo modo lo stimatore

$$\begin{aligned} \widehat{V}(\widehat{\mu}_q) &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in \mathbf{s}} (y_i - \widehat{R}x_i)^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \sum_{i \in \mathbf{s}} \widehat{e}_i^2 \end{aligned} \tag{6.11}$$

in cui le quantità $\widehat{e}_i = y_i - \widehat{R}x_i$ sono i residui introdotti in precedenza (vds. (6.4)). Per altri stimatori di $V(\widehat{\mu}_q)$, e per qualche sintetica nota sul loro confronto, si rinvia al volume di Cochran (1977), pp. 155–156.

Dalla (6.11), usando la solita approssimazione normale per la distribuzione di probabilità di $\widehat{\mu}_q$, si possono anche costruire intervalli di confidenza (approssimati) per la media della popolazione. Indicando con z_α il quantile di ordine α della distribuzione normale standard, è immediato verificare che

$$\left[\widehat{\mu}_q - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_q)},\ \widehat{\mu}_q + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_q)}\right]$$

è un intervallo di confidenza per la media μ_y della popolazione, al livello (approssimato) $1 - \alpha$.

Esempio 6.2. Consideriamo ancora l'Esempio 6.1. In questo caso, la stima della varianza di $\widehat{\mu}_q$ (6.11) assume il valore:

$$\widehat{V}(\widehat{\mu}_q) = \left(\frac{1}{24} - \frac{1}{426}\right) \frac{1}{23} \sum_{i=1}^{24} \widehat{e}_i^2 = 2.40.$$

Sulla base di questa stima, non è difficile costruire intervalli di confidenza per μ_y. A titolo di esempio, costruiamo un intervallo di confidenza per μ_y al livello 0.95. Essendo $z_{0.025} = 1.96$, si ha che

$$\left[107.56 - 1.96\sqrt{2.40},\ 107.56 + 1.96\sqrt{2.40}\right] = [104.52,\ 110.60]$$

è un intervallo di confidenza approssimato per μ_y, di livello 0.95. □

6.4 Stimatore di tipo media di rapporti*

Lo stimatore per quoziente (6.3) è essenzialmente basato su un *rapporto tra medie campionarie*. Una semplice idea alternativa per costruire uno stimatore

che sfrutti la conoscenza del carattere ausiliario $\mathcal{X}$ potrebbe essere quella di basarsi su una *media di rapporti* y_i/x_i. Per semplificare un po' la trattazione, definiamo sulle N unità della popolazione un nuovo carattere $\mathcal{Z}$ il quale, in corrispondenza dell'unità i, assume modalità $z_i = y_i/x_i$ $(i = 1, \ldots, N)$. Usando la consueta simbologia, siano

$$\mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i,\ \sigma_z^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \mu_z)^2,\ \sigma_{xz} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)\,(z_i - \mu_z)$$

la media e la varianza di $\mathcal{Z}$, e la covarianza tra $\mathcal{X}$ e $\mathcal{Z}$. Siano inoltre, come al solito,

$$S_z^2 == \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \mu_z)^2,\ S_{xz} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)\,(z_i - \mu_z).$$

L'idea di base per costruire uno stimatore di μ_y basato su una media di rapporti è quella di partire dalla media campionaria delle z_i, moltiplicata per μ_x:

$$t = \mu_x\,\overline{z}_{\mathbf{s}} = \mu_x\,\frac{1}{n}\sum_{i\in\mathbf{s}} z_i = \mu_x\,\frac{1}{n}\sum_{i\in\mathbf{s}}\frac{y_i}{x_i}. \tag{6.12}$$

Essendo il disegno usato di tipo ssr, il valore atteso di $\overline{z}_{\mathbf{s}}$ è uguale a μ_z, e quindi si può scrivere

$$E[t] = \mu_x E[\overline{z}_{\mathbf{s}}] = \mu_x\,\mu_z.$$

Tenendo conto che $y_i = x_i\,z_i$, ne consegue che la distorsione dello stimatore t (6.12) è pari a:

$$\begin{aligned} B(t) &= E[t] - \mu_y \\ &= \mu_x\,\mu_z - \mu_y \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} z_i\right) - \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= -\left\{\frac{1}{N}\sum_{i=1}^{N} x_i\,z_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} z_i\right)\right\} \\ &= -\sigma_{xz}. \end{aligned} \tag{6.13}$$

Ora, come noto (vds. in particolare la Sezione 3.7), uno stimatore corretto di σ_{xz} è il seguente:

$$\begin{aligned}\widehat{\sigma}_{xz} &= \frac{N-1}{N}\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})\,(z_i-\overline{z}_\mathbf{s})\\ &= \frac{N-1}{N}\frac{n}{n-1}\frac{1}{n}\sum_{i\in\mathbf{s}}(x_i-\overline{x}_\mathbf{s})\,(z_i-\overline{z}_\mathbf{s})\\ &= \frac{N-1}{N}\frac{n}{n-1}\left(\frac{1}{n}\sum_{i\in\mathbf{s}}x_i\,z_i-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s}\right)\\ &= \frac{N-1}{N}\frac{n}{n-1}\,(\overline{y}_\mathbf{s}-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s})\end{aligned}$$

per cui dalla (6.13) segue la relazione

$$E\left[\frac{N-1}{N}\frac{n}{n-1}\,(\overline{y}_\mathbf{s}-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s})\right]=\sigma_{xz}=-B(t). \tag{6.14}$$

La (6.14) è di fondamentale importanza per costruire uno stimatore corretto di μ_y a partire dalla media di rapporti t (6.12). Infatti, consideriamo il nuovo stimatore di μ_y:

$$\widehat{\mu}_{HR}=t+\frac{N-1}{N}\frac{n}{n-1}\,(\overline{y}_\mathbf{s}-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s})=\mu_x\,\overline{z}_\mathbf{s}+\frac{N-1}{N}\frac{n}{n-1}\,(\overline{y}_\mathbf{s}-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s}) \tag{6.15}$$

noto in letteratura come *stimatore di Hartley-Ross*. La sua più importante proprietà è la correttezza:

$$\begin{aligned}E\left[\widehat{\mu}_{HR}\right] &= \mu_x\,E\left[\overline{z}_\mathbf{s}\right]+E\left[\frac{N-1}{N}\frac{n}{n-1}\,(\overline{y}_\mathbf{s}-\overline{x}_\mathbf{s}\,\overline{z}_\mathbf{s})\right]\\ &= \mu_x\,\mu_z+\sigma_{xz}\\ &= \mu_y-\sigma_{xz}+\sigma_{xz}\\ &= \mu_y.\end{aligned}$$

La costruzione dello stimatore (6.15) sottende un importante principio, che è bene mettere nel dovuto rilievo. Si consideri un generico stimatore $\widehat{\theta}$ di un parametro θ, e si supponga che $B(\widehat{\theta})=E[\widehat{\theta}]-\theta$ sia la sua distorsione. Si supponga poi di essere in grado di costruire uno stimatore corretto $\widehat{B}$ della distorsione di $\widehat{\theta}$: $E[\widehat{B}]=B(\widehat{\theta})$. Allora, il nuovo stimatore $\widetilde{\theta}=\widehat{\theta}-\widehat{B}$ è uno stimatore corretto di θ. In sostanza, questo è quel che è stato fatto nella presente sezione, in cui t svolge il ruolo dello "stimatore iniziale" $\widehat{\theta}$, e $-\widehat{\sigma}_{xz}$ è uno stimatore corretto della distorsione di t. Un'applicazione di questo principio è presentata nell'Esercizio 6.4.

Esercizi

6.1. Provare la relazione (6.5).

Suggerimento. $B(\widehat{\mu}_q) = -(E[\widehat{R}\,\overline{x}_{\mathbf{s}}] - E[\widehat{R}]\,E[\overline{x}_{\mathbf{s}}])$.

6.2. Provare che vale la disuguaglianza $|B(\widehat{\mu}_q)| \leqslant \sqrt{V(\widehat{R})\,V(\overline{x}_{\mathbf{s}})}$.

Suggerimento. Usare la (6.5) e la disuguaglianza di Schwarz.

6.3. Si consideri una popolazione di $N = 5$ unità, sulle quali sono definiti un carattere di interesse $\mathcal{Y}$ ed un carattere ausiliario $\mathcal{X}$. Le modalità dei due caratteri sono qui sotto riportate.

Unità i	1	2	3	4	5
Valori x_i	2	6	2	3	5
Valori y_i	10	31	11	14	23

a. Enumerare tutti i campioni ssr di $n = 3$ unità della popolazione, e per ognuno di essi calcolare il valore assunto dallo stimatore per quoziente $\widehat{\mu}_q$.
b. Calcolare esattamente il valore atteso, la varianza e l'errore quadratico medio dello stimatore $\widehat{\mu}_q$.
c. Confrontare l'errore quadratico medio dello stimatore $\widehat{\mu}_q$ con quello della media campionaria $\overline{x}_{\mathbf{s}}$.

6.4. Lo *stimatore per prodotto* di μ_y è definito come

$$\widehat{\mu}_p = \frac{\overline{x}_{\mathbf{s}}\,\overline{y}_{\mathbf{s}}}{\mu_x}.$$

a. Verificare che $\widehat{\mu}_p$ è uno stimatore distorto di μ_y, con distorsione

$$B(\widehat{\mu}_p) = \frac{1}{\mu_x}\left(\frac{1}{n} - \frac{1}{N}\right) S_{xy}.$$

b. Provare che

$$\widehat{\mu}_{pc} = \widehat{\mu}_p - \frac{1}{\mu_x}\left(\frac{1}{n} - \frac{1}{N}\right)\widehat{s}_{xy}$$

è uno stimatore corretto di μ_y.

6.5. Si consideri la popolazione del *file* `cultura.txt`, in cui sono riportate (oltre ad altri dati) le spese annue per attività culturali di $N = 1500$ famiglie (carattere $\mathcal{Y}$ di interesse) e il numero di componenti di ciascuna famiglia (carattere ausiliario $\mathcal{X}$, da considerare noto *a priori*). L'obiettivo è di stimare la spesa media annua μ_y sostenuta dalle famiglie per attività culturali.

a. Selezionare un campione ssr di $n = 60$ famiglie della popolazione, di cui si osservano numero di componenti e spese per attività culturali.
b. Stimare μ_y tramite lo stimatore per quoziente.
c. Stimare la varianza dello stimatore per quoziente costruito al punto *b*.
d. Costruire un intervallo di confidenza (approssimato) per μ_y al livello 0.96.

7

Disegno campionario stratificato I

7.1 Motivazioni e aspetti di base

La costruzione degli stimatori di regressione e quoziente si basa sulla disponibilità di *informazioni ausiliarie* sulla popolazione oggetto di studio, ed in particolare sulla conoscenza di un *carattere ausiliario*. Questi due stimatori sono stati studiati rispettivamente nei Capitoli 5, 6, fermo restando il tipo di disegno campionario usato, di tipo ssr. Questo, però, non è l'unico modo di sfruttare informazioni ausiliarie. Come già accennato all'inizio del Capitolo 5, le informazioni ausiliarie possono anche essere usate, in alcuni casi, per costruire un disegno campionario che non sia quello ssr. È questo, per l'appunto, il caso del disegno stratificato. Per comprendere meglio le idee di base del disegno stratificato, iniziamo con un semplice esempio.

Esempio 7.1. Nel *file* `stature.txt`, già usato più volte nel Capitolo 3, sono riportati numeri di matricola, sesso e statura di una popolazione di $N = 1570$ studenti universitari. La statura media della popolazione è pari a $\mu_y = 172.80$, e la varianza a $\sigma_y^2 = 59.9$. Nella popolazione vi sono in totale 750 femmine e 820 maschi. La statura media delle femmine e quella dei maschi sono rispettivamente pari a

$$\mu_{y\,fem} = 168.26, \quad \mu_{y\,mas} = 177.00$$

e tra queste e la media μ_y intercorre la relazione (cfr. Sezione 1.4):

$$\mu_y = \frac{750}{1570}\,\mu_{y\,fem} + \frac{820}{1570}\,\mu_{y\,mas}. \tag{7.1}$$

Per stimare μ_y selezioniamo un campione di $n = 100$ unità mediante disegno ssr, e calcoliamo la corrispondente media campionaria. Il campione selezionato è contenuto nel *file* `campstature_ssr.txt`, e la media campionaria corrispondente è $\overline{y}_{\mathbf{s}} = 171.98$.

Come già messo in evidenza, l'efficienza, la precisione della media campionaria, quando usata con il disegno ssr, dipende da due elementi: la numero-

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_7,

sità n del campione e la varianza σ_y^2 della popolazione. La media campionaria è tanto meno precisa quanto più elevata è la varianza della popolazione.

Nel nostro caso, la variabilità delle stature che si osserva nella popolazione di studenti dipende da *due* fattori: (a) la diversità di stature di studenti dello stesso sesso; (b) la diversità di stature di studenti di sesso diverso. In effetti, la statura degli studenti maschi è mediamente superiore a quella delle femmine, e questo comporta una conseguenza rilevante: vi è molta più omogeneità tra le stature di studenti dello stesso sesso che tra quelle di studenti della popolazione complessiva. Pertanto, all'imprecisione della media campionaria contribuisce il fatto che si campiona dalla popolazione totale dei 1570 studenti, in cui si trovano sia maschi che femmine e con stature assai disomogenee.

Un'idea molto naturale per conseguire una maggior precisione di stima, a parità di numerosità campionaria, potrebbe essere quella di estrarre separatamente un campione di maschi ed uno di femmine, e nello stimare da un lato la statura media degli studenti maschi, e dall'altro quella delle femmine. Questo, ovviamente, richiede che si disponga di una lista dei soli studenti maschi, e di una dei soli studenti femmine da cui selezionare i due campioni. Per rendere confrontabili i risultati con quanto ottenuto usando il disegno ssr si devono selezionare, in totale, 100 unità. Per il momento non disponiamo di linee guida per scegliere quanti studenti maschi e quante femmine campionare, per cui scegliamo, arbitrariamente, di selezionare:

- un campione ssr di $n = 48$ delle 750 unità della sottopopolazione degli studenti femmine;
- un campione ssr di $n = 52$ delle 820 unità della sottopopolazione degli studenti maschi.

I relativi dati campionari sono contenuti, rispettivamente, nei *file* `campione_f.txt` e `campione_m.txt`. Le corrispondenti medie campionarie sono rispettivamente uguali a

$$\overline{y}_{\mathbf{s}\,fem} = 168.29, \;\; \overline{y}_{\mathbf{s}\,mas} = 176.98 \tag{7.2}$$

e costituiscono stime rispettivamente di $\mu_{y\,fem}$ e $\mu_{y\,mas}$.

Il problema è ora quello di combinare le due stime (7.2) per ottenere una stima di μ_y. Un'idea molto semplice è quella di fare riferimento alla (7.1), con $\mu_{y\,fem}$ e $\mu_{y\,mas}$ rimpiazzati rispettivamente da $\overline{y}_{\mathbf{s}\,fem}$ e $\overline{y}_{\mathbf{s}\,mas}$. Si ottiene in questo modo la stima:

$$\frac{750}{1570}\,\overline{y}_{\mathbf{s}\,fem} + \frac{820}{1570}\,\overline{y}_{\mathbf{s}\,mas} = 172.80.$$

La stima ora ottenuta è più precisa della media campionaria, a parità di numero di unità campionarie. Questo, come sopra accennato, è dovuto al fatto che le sottopopolazioni degli studenti dello stesso sesso sono molto più omogenee di quella totale di tutti gli studenti, e quindi si possono ottenere stime precise di $\mu_{y\,fem}$ e $\mu_{y\,mas}$ anche con poche unità campionarie. □

Supponiamo che la popolazione totale I_N sia divisa in M sottopopolazioni, o *strati* , rispettivamente di numerosità $N_1, N_2, \ldots, N_M$, con $N_1 + N_2 + \cdots + N_M = N$. I principî di base sono già stati introdotti nella Sezione 1.4. Indichiamo con $I^1_{N_1}, I^2_{N_2}, \ldots, I^M_{N_M}$ gli M strati, e sia $w_g = N_g/N$ il *peso* dello strato g-mo ($g = 1, \ldots, M$). Chiaramente, valgono le relazioni

$$0 \leqslant w_g \leqslant 1 \; per \; ogni \; g = 1, \ldots, M; \; \sum_{g=1}^{M} w_g = 1.$$

Ogni unità della popolazione è ora identificata da una doppia etichetta (g, i), in cui:

- g $(= 1, \ldots, M)$ indica lo strato a cui appartiene l'unità;
- i $(= 1, \ldots, N_g)$ identifica l'unità nell'ambito dello strato di appartenenza.

Sempre in conformità con quanto detto nella Sezione 1.4, indicheremo con y_{gi} la modalità dell'unità i $(= 1, \ldots, N_g)$ dello strato g-mo ($g = 1, \ldots, M$), e con

$$\mu_{yg} = \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi}, \;\; \sigma^2_{yg} = \frac{1}{N_g} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2 \,; \; g = 1, \ldots, M$$

rispettivamente la media e la varianza del carattere di interesse $\mathcal{Y}$ nello strato g-mo. Come mostrato nella Proposizione 1.1, valgono le due seguenti, fondamentali relazioni

$$\mu_y = \sum_{g=1}^{M} w_g \, \mu_{yg}, \;\; \sigma^2_y = \sum_{g=1}^{M} w_g \, \sigma^2_{yg} + \sum_{g=1}^{M} w_g \, (\mu_{yg} - \mu_y)^2. \tag{7.3}$$

L'idea di base del disegno campionario stratificato è elementare. Esso consiste nel selezionare, mediante disegno ssr e indipendentemente da uno strato all'altro,

- un campione ssr $\mathbf{s}_1$ di numerosità n_1 dallo strato 1;
- un campione ssr $\mathbf{s}_2$ di numerosità n_2 dallo strato 2;
  ...
- un campione ssr $\mathbf{s}_M$ di numerosità n_M dallo strato M.

Il "campione totale" $\mathbf{s}$ è formato dagli M "sottocampioni" $\mathbf{s}_1, \ldots, \mathbf{s}_M$, ciascuno relativo ad uno degli strati. In simboli:

$$\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M) \tag{7.4}$$

e la sua numerosità è $n = n_1 + n_2 + \cdots + n_M$.

Del disegno stratificato non è difficile dare una descrizione formale specificando esattamente sia lo spazio dei campioni che le probabilità dei vari campioni.

La (7.4) mette in evidenza che in un campione stratificato le prime n_1 unità provengono dal primo strato, le successive n_2 unità provengono dal secondo strato, e così via fino alle ultime n_M unità, che provengono dallo strato M-mo. Ogni campione stratificato si può vedere quindi come una combinazione di classe n delle unità della popolazione, con il vincolo che n_1 unità provengano dallo strato $I^1_{N_1}$, ..., n_M provengano dallo strato $I^M_{N_M}$. Lo spazio dei campioni $\mathcal{S}$ del disegno stratificato è quindi l'insieme $\mathcal{C}_{(N_1,\ldots,N_M);(n_1,\ldots,n_M)}$ di tutte queste combinazioni. In simboli:

$$\begin{aligned}\mathcal{S} &= \mathcal{C}_{(N_1,\ldots,N_M);(n_1,\ldots,n_M)} \\ &= \mathcal{C}_{N_1,n_1} \times \cdots \times \mathcal{C}_{N_M,n_M} \\ &= \bigtimes_{g=1}^{M} \mathcal{C}_{N_g,n_g} \,.\end{aligned}$$

In secondo luogo, poiché il generico $\mathbf{s}_g$ è un campione ssr di n_g unità di $I^g_{N_g}$, si avrà

$$p(\mathbf{s}_g) = \frac{1}{\binom{N_g}{n_g}} \; \textit{per ogni } \mathbf{s}_g \in \mathcal{C}_{N_g,n_g};\; g = 1, \ldots, M.$$

Essendo inoltre gli M sottocampioni $\mathbf{s}_1$, ..., $\mathbf{s}_M$ indipendenti, si conclude che

$$\begin{aligned}p(\mathbf{s}) &= p(\mathbf{s}_1)\, p(\mathbf{s}_2) \cdots p(\mathbf{s}_M) \\ &= \frac{1}{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_M}{n_M}}\end{aligned}$$

per ogni campione $\mathbf{s}$ dello spazio $\mathcal{S}$ dei campioni.

7.2 Stima della media di una popolazione

La costruzione di uno stimatore della media μ_y si basa su considerazioni molto semplici, simili in linea di principio a quelle dell'Esempio 7.1. Il sottocampione $\mathbf{s}_g$ è un campione ssr di numerosità n_g dello strato $I^g_{N_g}$, $g = 1, \ldots, M$. Pertanto, come stimatore della media μ_{yg} dello strato stesso, si può utilizzare la media campionaria:

$$\overline{y}_g = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi}.$$

Si hanno in tal modo M stime $\overline{y}_1$, ..., $\overline{y}_M$, una per la media di ogni strato. Tali stime devono poi essere ricombinate per produrre una stima della media μ_y dell'intera popolazione. Per ricombinarle usiamo la prima delle (7.3). Si ottiene in questo modo lo stimatore:

$$\widehat{\mu}_{str} = \sum_{g=1}^{M} w_g\, \overline{y}_g. \tag{7.5}$$

Le proprietà dello stimatore (7.5) sono studiate nella Proposizione 7.1. Per comodità di notazione, indicheremo con

$$S_{yg}^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2 = \frac{N_g}{N_g - 1} \sigma_{yg}^2 \,;\; g = 1, \ldots, M$$

la varianza corretta dello strato g-mo ($g = 1, \ldots, M$).

Proposizione 7.1. *Se il disegno campionario è stratificato, $\widehat{\mu}_{str}$ è uno stimatore corretto della media della popolazione:*

$$E\left[\widehat{\mu}_{str}\right] = \mu_y \tag{7.6}$$

e la sua varianza è pari a

$$V\left(\widehat{\mu}_{str}\right) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2. \tag{7.7}$$

Dimostrazione. Il sottocampione $\mathbf{s}_g$ è selezionato dallo strato g-mo mediante disegno ssr, per cui, usando risultati noti, si ha

$$E\left[\overline{y}_g\right] = \mu_{yg}, \;\; V\left(\overline{y}_g\right) = \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2; \;\; g = 1, \ldots, M. \tag{7.8}$$

Usando la prima delle (7.8) si prova la correttezza di $(\widehat{\mu}_{str})$:

$$\begin{aligned} E\left[\widehat{\mu}_{str}\right] &= E\left[\sum_{g=1}^{M} w_g \, \overline{y}_g\right] \\ &= \sum_{g=1}^{M} w_g \, E\left[\overline{y}_g\right] \\ &= \sum_{g=1}^{M} w_g \, \mu_{yg} \\ &= \mu_y. \end{aligned}$$

Per quanto riguarda la varianza, essendo i sottocampioni $\mathbf{s}_1, \ldots, \mathbf{s}_M$ indipendenti, anche le medie campionarie $\overline{y}_1, \ldots, \overline{y}_M$ sono indipendenti, e quindi le loro covarianze sono nulle. Ne consegue che

$$\begin{aligned} V\left(\widehat{\mu}_{str}\right) &= V\left(\sum_{g=1}^{M} w_g \, \overline{y}_g\right) \\ &= \sum_{g=1}^{M} w_g^2 \, V\left(\overline{y}_g\right) \\ &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 \end{aligned}$$

ossia la (7.7). □

Usando risultati noti per il disegno ssr, è anche facile costruire uno stimatore corretto della varianza di $\widehat{\mu}_{str}$. Sia

$$\widehat{s}^2_{yg} = \frac{1}{n_g - 1} \sum_{i \in \mathbf{s}_g} (y_{gi} - \overline{y}_g)^2$$

la varianza campionaria corretta dello strato g-mo ($g = 1, \ldots, M$). Come conseguenza della Proposizione 3.3, si ha

$$E\left[\widehat{s}^2_{yg}\right] = S^2_{yg}; \;\; g = 1, \ldots, M$$

e quindi è immediato provare la seguente proposizione.

Proposizione 7.2. *Se il disegno campionario è stratificato, lo stimatore*

$$\widehat{V}(\widehat{\mu}_{str}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \widehat{s}^2_{yg} \tag{7.9}$$

è uno stimatore corretto di $V(\widehat{\mu}_{str})$:

$$E\left[\widehat{V}(\widehat{\mu}_{str})\right] = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S^2_{yg} = V(\widehat{\mu}_{str}).$$

La costruzione di intervalli di confidenza, infine, procede lungo linee simili a quelle sviluppate nei capitoli precedenti. Se le numerosità campionarie di strato, $n_1, \ldots, n_M$ sono sufficientemente grandi, le medie campionarie di strato $\overline{y}_1, \ldots, \overline{y}_M$ avranno distribuzione approssimativamente normale. Ne consegue che anche $\widehat{\mu}_{str} = \sum_{g=1}^{M} w_g \overline{y}_g$ ha distribuzione approssimata di tipo normale, con media μ_y e varianza $V(\widehat{\mu}_{str})$. Ragionando esattamente come nei capitoli precedenti, e sostituendo l'incognita $V(\widehat{\mu}_{str})$ con la sua stima (7.9), si ha che

$$\frac{\widehat{\mu}_{str} - \mu_y}{\sqrt{\widehat{V}(\widehat{\mu}_{str})}}$$

ha distribuzione approssimata di tipo normale standard. Detto pertanto, come al solito, z_α il quantile di ordine α della distribuzione normale standard, è immediato verificare che

$$\left[\widehat{\mu}_{str} - z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_{str})},\; \widehat{\mu}_{str} + z_{\alpha/2}\sqrt{\widehat{V}(\widehat{\mu}_{str})}\right]$$

è un intervallo di confidenza per μ_y, al livello approssimato $1 - \alpha$.

Esempio 7.2. Si consideri ancora la popolazione di 1570 unità del *file* `stature.txt`. Si è già visto che $\widehat{\mu}_{str} = 172.80$. Le varianze campionarie (corrette) di strato sono uguali (con la notazione già usata nell'Esempio 7.1) a

$$S^2_{y\,fem} = 39.25,\; S^2_{y\,mas} = 54.33\,,$$

per cui lo stimatore (7.9) risulta pari a

$$\widehat{V}(\widehat{\mu}_{str}) = \left(\frac{750}{1570}\right)^2 \left(\frac{1}{48} - \frac{1}{750}\right) \widehat{s}^2_{y\,fem} + \left(\frac{820}{1570}\right)^2 \left(\frac{1}{52} - \frac{1}{820}\right) \widehat{s}^2_{y\,mas} = 0.44.$$

Tenendo infine conto che $z_{0.005} = 2.576$, un intervallo di confidenza approssimato per μ_y al livello $1 - \alpha = 0.99$ è il seguente

$$\left[172.80 - 2.576\sqrt{0.44},\ 172.80 + 2.576\sqrt{0.44}\right] = [171.09,\ 174.51]\,. \qquad \square$$

La teoria di base del campionamento stratificato è molto semplice. In effetti, si tratta poco più che di un'applicazione di concetti già ampiamente illustrati a proposito del disegno ssr. Vi sono però alcuni importanti punti da trattare, che verranno studiati in questo e nel successivo capitolo, e che sono qui di seguito brevemente elencati.

- *Allocazione delle osservazioni agli strati.* Data la numerosità campionaria totale n, quante unità bisogna selezionare da ciascuno strato? In altre parole, in che modo scegliere $n_1, \ldots, n_M$ una volta che sia dato n? È questo il problema dell'*allocazione* delle unità campionarie ai vari strati. Detta $a_g = n_g/n$ la proporzione di unità campionarie allocate allo strato g-mo, si tratta in sostanza di stabilire i valori da assegnare a $a_1, \ldots, a_M$ (notare che le a_g sono tutte non negative, e che $a_1 + \cdots + a_M = 1$).
- *Scelta della numerosità campionaria n.*
- *Definizione degli strati.* La definizione effettiva degli strati da impiegare richiede di decidere sia il *numero* degli strati, sia di stabilire qualche criterio per la loro *costruzione*.

7.3 Campionamento stratificato proporzionale

Il campionamento stratificato proporzionale è il più semplice tra i disegni campionari di tipo stratificato. Esso prevede che le numerosità campionarie di strato siano *proporzionali ai pesi degli strati*:

$$n_g = n\, w_g, \quad g = 1, \ldots, M. \tag{7.10}$$

In un certo senso, l'idea su cui si basa la regola di allocazione proporzionale (7.10) è di costruire una "versione ridotta" della popolazione. Infatti, la (7.10) equivale a

$$\frac{n_g}{n} = \frac{N_g}{N}; \quad g = 1, \ldots, M$$

da cui si evince che le proporzioni delle numerosità campionarie dei diversi strati rispetto alla numerosità campionaria totale sono uguali alle corrispondenti proporzioni a livello di popolazione. In simboli: $a_g = w_g$, $g = 1, \ldots, M$.

Nel caso speciale del disegno stratificato proporzionale lo stimatore (7.5) assume una forma particolarmente semplice, in quanto si riduce alla media campionaria. Per mostrare questo fatto basta osservare che, per la (7.10),

$$\begin{aligned}\widehat{\mu}_{str} &= \sum_{g=1}^{M} w_g \, \overline{y}_g = \sum_{g=1}^{M} \frac{n_g}{n} \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi} = \frac{1}{n} \sum_{g=1}^{M} \sum_{i \in \mathbf{s}_g} y_{gi} \\ &= \overline{y}_{\mathbf{s}}.\end{aligned}$$

Anche la varianza dello stimatore stimatore (7.5), nel caso di allocazione proporzionale, assume una forma semplice. Si ha infatti, usando la (7.7) e tenendo conto che $w_g = N_g/N = n_g/n$,

$$\begin{aligned}V\left(\widehat{\mu}_{str};\ prop\right) &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{g=1}^{M} w_g S_{yg}^2. \end{aligned} \tag{7.11}$$

Lo scopo della stratificazione, come più volte asserito, è essenzialmente quello di produrre un'efficienza di stima superiore a quella che si ottiene con il disegno semplice. Per una data numerosità campionaria totale n, l'efficienza dello stimatore $\widehat{\mu}_{str}$ dipende dalle numerosità campionarie di strato $n_1, \ldots, n_M$. In generale non è affatto detto che, a parità di numerosità totale n del campione, un qualsiasi disegno stratificato produca risultati migliori del disegno ssr. Un caso in cui ciò frequentemente accade è proprio quello del disegno proporzionale. Per precisare questa affermazione è necessario confrontare, a parità di numerosità campionaria n, la varianza di $\widehat{\mu}_{str}$, quando usato con il disegno stratificato proporzionale, con la varianza della media campionaria quando usata con il disegno ssr. Utilizzando i risultati della Sezione 1.4 è facile verificare (Esercizio 7.1) che vale la relazione

$$\begin{aligned}&V(\overline{y}_{\mathbf{s}};\ ssr) = V\left(\widehat{\mu}_{str};\ prop\right) \\ &+ \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N}{N-1} \left\{ \sum_{g=1}^{M} w_g (\mu_{yg} - \mu_y)^2 - \frac{1}{N} \sum_{g=1}^{M} (1 - w_g) \, S_{yg}^2 \right\}. \end{aligned} \tag{7.12}$$

Pertanto, il disegno stratificato proporzionale fornisce, in termini di efficienza di stima, risultati migliori del disegno ssr se e solo se

$$\sum_{g=1}^{M} w_g (\mu_{yg} - \mu_y)^2 - \frac{1}{N} \sum_{g=1}^{M} (1 - w_g) \, S_{yg}^2 > 0$$

ovvero se e solo se

$$\sum_{g=1}^{M} w_g (\mu_{yg} - \mu_y)^2 > \frac{1}{N} \sum_{g=1}^{M} (1 - w_g) \, S_{yg}^2. \tag{7.13}$$

La (7.13) vale in genere se le numerosità di strato N_g sono "grandi", così che tale è anche la numerosità totale N della popolazione. In questo caso il termine $1/N$ assume valori molto piccoli, prossimi a zero, per cui, a meno di casi eccezionali, anche il termine $\sum_g (1 - w_g)\, S^2_{yg}/N$ assumerà un valore piccolo, tipicamente minore di $\sum_g w_g(\mu_{yg} - \mu_y)^2$. Ora, nella stragrande maggioranza dei casi che si incontrano nelle applicazioni le numerosità campionarie di strato sono abbastanza grandi da giustificare l'assunzione $\sum_g (1 - w_g)\, S^2_{yg}/N \approx 0$, per cui in genere il disegno stratificato proporzionale fornisce risultati migliori del disegno ssr a parità di numerosità campionaria.

Esempio 7.3. Non tutti i disegni stratificati, come detto, producono stime più efficienti di quello ssr. Mostriamo questo fatto con un semplice esempio. Si consideri una popolazione di $N = 2000$ unità, con varianza corretta $S^2_y = 2600$. Se si seleziona un campione ssr di numerosità $n = 90$, e si usa la media campionaria $\overline{y}_{\mathbf{s}}$ per stimare la media μ_y della popolazione, la sua varianza sarà pari (con ovvia notazione) a

$$V(\overline{y}_{\mathbf{s}};\ ssr) = \left(\frac{1}{90} - \frac{1}{2000}\right) 2600 = 27.59.$$

Supponiamo ora che la popolazione sia suddivisa in tre strati, rispettivamente di numerosità $N_1 = 1400$, $N_2 = 400$, $N_3 = 200$. I pesi di strato sono $w_1 = 0.7$, $w_2 = 0.2$, $w_3 = 0.1$. Supponiamo anche che le varianze (corrette) di strato siano pari rispettivamente a $S^2_{y1} = 1500$, $S^2_{y2} = 4000$, $S^2_{y3} = 5000$, da cui $\sum_g w_g\, S^2_{yg} = 2350$.

Se dalla popolazione si estrae un campione stratificato proporzionale di numerosità totale $n = 90$, così che $n_1 = n\, w_1 = 63$, $n_2 = n\, w_2 = 18$, $n_3 = n\, w_3 = 9$, si ha

$$V(\widehat{\mu}_{str};\ prop) = \left(\frac{1}{90} - \frac{1}{2000}\right) 2350 = 24.94.$$

Se invece dalla popolazione si estrae un campione stratificato sempre di numerosità totale $n = 90$, ma che seleziona da ogni strato lo stesso numero di unità: $n_1 = n_2 = n_3 = 30$ (allocazione uniforme) , si ha

$$\begin{aligned} V(\widehat{\mu}_{str};\ unif) &= \sum_{g=1}^{3} \left(\frac{1}{n_g} - \frac{1}{N_g}\right) w_g^2 S^2_{yg} \\ &= \frac{1}{30} \sum_{g=1}^{3} w_g^2 S^2_{yg} - \frac{1}{2000} \sum_{g=1}^{3} w_g\, S^2_{yg} \\ &= \frac{945}{30} - \frac{2350}{2000} \\ &= 30.325. \end{aligned}$$

A parità di numerosità campionaria, in questo caso si ha $V(\widehat{\mu}_{str};\ prop) < V(\overline{y}_{\mathbf{s}};\ ssr) < V(\widehat{\mu}_{str};\ unif)$. □

7.3.1 L'effetto del disegno

I risultati sopra ottenuti possono essere espressi in termini di effetto del disegno, introdotto nel Capitolo 3. Si consideri un disegno stratificato di numerosità totale n. La varianza dello stimatore $\widehat{\mu}_{str}$, data dalla (7.7), dipende dall'allocazione del campione negli M strati. Se l'allocazione è proporzionale, allora $\widehat{\mu}_{str} = \overline{y}_{\mathbf{s}}$ e la varianza (7.7) diventa pari alla (7.11). L'effetto del disegno è quindi pari a

$$\begin{aligned} Deff\,(str\ prop,\ \overline{y}_{\mathbf{s}}) &= \frac{V(\overline{y}_{\mathbf{s}};\ prop)}{V(\overline{y}_{\mathbf{s}};\ ssr)} \\ &\approx \frac{\sum_{g=1}^{M} w_g S_{yg}^2}{\sum_{g=1}^{M} W_g (S_{yg}^2 + (\mu_{yg} - \mu_y)^2)} \end{aligned} \tag{7.14}$$

dove si è usata l'approssimazione $\sum_g (1 - w_g)\, S_{yg}^2 / N \approx 0$. A meno che le medie di strato non siano tutte uguali (e quindi uguali alla media della popolazione), l'effetto del disegno sarà sempre minore di 1. Ciò significa che la stratificazione proporzionale implica un guadagno di precisione rispetto a un campionamento casuale semplice. Tale guadagno risulterà tanto più elevato quanto più elevata è la differenza tra le medie di strato. Se ad esempio $Deff\,(str\ prop,\ \overline{y}_{\mathbf{s}}) = 0.7$, si ha una riduzione della varianza del 30% rispetto ad un campionamento casuale semplice. Di conseguenza, la dimensione campionaria efficace risulta pari a

$$n_{eff}\,(str\ prop,\ \overline{y}_{\mathbf{s}}) = \frac{n}{0.7} = 1.43 \times n.$$

Questo significa che se si utilizzasse un campionamento casuale semplice al posto di un campionamento stratificato proporzionale, sarebbe necessario estrarre un campione di $1.43 \times n$ unità per ottenere la stessa precisione.

L'allocazione proporzionale garantisce una precisione almeno pari a quella del campionamento semplice senza ripetizione, come sottolineato dalla (7.14). Spesso, però, nelle rilevazioni campionarie si impone la necessità di ricorrere a un campionamento stratificato con allocazione non proporzionale. Per esempio, poiché i costi di raccolta delle informazioni possono differire da strato a strato, gli strati con un maggior costo di rilevazione saranno caratterizzati da una frazione di campionamento inferiore a quella che si avrebbe nel caso proporzionale. Questo punto sarà esaminato nella sezione successiva.

Spesso, poi, tra gli obiettivi di indagini campionarie vi è quello di fornire stime caratterizzate da un certo livello di precisione per particolari gruppi di unità (sottopopolazioni) denominati *domini di studio*. Tali esigenze di precisione portano spesso ad abbandonare l'allocazione proporzionale.

Situazione siffatte implicano che in alcuni strati è necessario allocare più unità di quelle previste da una allocazione proporzionale, e in altri meno unità. Il guadagno in precisione che caratterizza un'allocazione proporzionale non necessariamente si verifica in un'allocazione non proporzionale, come evidenziato nell'Es. 7.3 e nel successivo Es. 7.4.

Esempio 7.4. Supponiamo per semplicità che $S^2_{yg} = S^2_y$ per ciascuno strato $g = 1, \ldots, M$, e che il fattore di correzione per popolazioni finite sia trascurabile: $(1 - n_g/N_g) \approx 1$. L'ipotesi $S^2_{yg} = S^2$ implica che:

- le varianze di strato siano costanti;
- le medie di strato siano approssimativamente uguali $\mu_{yg} = \mu_y$ per ogni $g = 1, \ldots, M$.

In tali condizioni l'effetto del disegno assume la forma

$$\begin{aligned} \mathit{Deff}(str, \widehat{\mu}_{str}) &= \frac{V(\widehat{\mu}_{str};\ str)}{V(\overline{y}_{\mathbf{s}};\ ssr)} \\ &\approx n \sum_{g=1}^{M} \frac{W_g^2}{n_g} \end{aligned} \tag{7.15}$$

e questa quantità non è necessariamente minore di 1. Si considerino ad esempio due strati della popolazione che contengono rispettivamente il 70% e il 30% delle unità della popolazione stessa. Formalmente $w_1 = 2.33 \times w_2$, con $w_2 = 0.3$. Allo scopo di ottenere stime delle medie dei due strati caratterizzate dallo stesso livello di precisione supponiamo di estrarre campioni di eguale numerosità: $n_1 = n_2 = 1000$. Applicando la (7.15) si ricava un effetto del disegno $\mathit{Deff}(str, \widehat{\mu}_{str}) = 1.16$ e una dimensione campionaria efficace pari a

$$n_{\mathit{eff}}(str, \widehat{\mu}_{str}) = 2000/1.16 = 1724.$$

La richiesta di ottenere stime caratterizzate dallo stesso livello di precisione in ogni strato impone la selezione di un campione totale pari a $n_1 + n_2 = 2000$, mentre a livello di popolazione sarebbe stato possibile ottenere la stessa precisione con un campionamento casuale semplice di 1724 unità. □

7.4 Disegno stratificato ottimale

7.4.1 Allocazione di Neyman

Il disegno proporzionale fornisce una semplice regola per scegliere le numerosità campionarie $n_1, \ldots, n_M$ dei singoli strati una volta fissata la numerosità campionaria totale n. Ora, questa non è la sola regola di allocazione delle unità campionarie ai diversi strati. Un criterio di allocazione molto intuitivo potrebbe essere quello di scegliere $n_1, \ldots, n_M$ (sempre fissato n) in modo da rendere massima l'efficienza di stima della media della popolazione, ossia in modo da minimizzare la varianza dello stimatore $\widehat{\mu}_{str}$.

Sulla base della (7.7), la varianza di $\widehat{\mu}_{str}$ si può scrivere come

$$V(\widehat{\mu}_{str}) = \sum_{g=1}^{M} w_g^2 \frac{1}{n_g} S^2_{yg} - \sum_{g=1}^{M} w_g^2 \frac{1}{N_g} S^2_{yg} = \sum_{g=1}^{M} \frac{w_g^2 S^2_{yg}}{n_g} - \frac{1}{N} \sum_{g=1}^{M} w_g S^2_{yg}$$

in cui il termine $\sum_g w_g S_{yg}^2/N$ non dipende da $n_1, \ldots, n_M$. Pertanto, minimizzare rispetto a $n_1, \ldots, n_M$ la $V(\widehat{\mu}_{str})$ equivale a minimizzare il solo termine $\sum_g w_g^2 S_{yg}^2/n_g$. La conseguenza di tutto questo è che il problema di minimizzare $V(\widehat{\mu}_{str})$, fissato $n = n_1 + \cdots + n_M$, si può riscrivere come problema di minimo vincolato nel modo seguente:

$$\begin{cases} \textit{minimizzare}: \ \sum_{g=1}^{M} \frac{w_g^2 S_{yg}^2}{n_g} \\ \textit{con il vincolo}: \ \sum_{g=1}^{M} n_g = n \end{cases} . \tag{7.16}$$

Proposizione 7.3. *La soluzione del problema di ottimo* (7.16) *è del tipo:*

$$n_g = n \frac{w_g S_{yg}}{\sum_{h=1}^{M} w_h S_{yh}}; \ g = 1, \ldots, M. \tag{7.17}$$

Dimostrazione. Per risolvere il problema (7.16) si può usare la tecnica dei moltiplicatori di Lagrange. La funzione Lagrangiana assume la forma

$$\mathcal{L}(n_1, \ldots, n_M \ \lambda) = \sum_{g=1}^{M} \frac{w_g^2 S_{yg}^2}{n_g} + \lambda \left(\sum_{g=1}^{M} n_g - n \right) \tag{7.18}$$

dove λ è il moltiplicatore di Lagrange. Derivando la (7.18) rispetto a $n_1, \ldots, n_M$, λ e annullando le derivate si ottengono le $M + 1$ equazioni

$$\frac{\partial \mathcal{L}}{\partial n_g} = -\frac{w_g^2 S_{yg}^2}{n_g^2} + \lambda = 0 \ \ \textit{da cui segue che}$$

$$n_g = \frac{1}{\sqrt{\lambda}} w_g S_{yg}; \ g = 1, \ldots, M \tag{7.19}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{g=1}^{M} n_g - n = 0 \ \ \textit{da cui segue che} \ \ \sum_{g=1}^{M} n_g = n. \tag{7.20}$$

Sommando membro a membro le (7.19) e sfruttando la (7.20) si ha poi la relazione

$$\frac{1}{\sqrt{\lambda}} \sum_{g=1}^{M} w_g S_{yg} = \sum_{g=1}^{M} n_g = n$$

da cui si desume che

$$\sqrt{\lambda} = \frac{1}{n} \sum_{g=1}^{M} w_g S_{yg}. \tag{7.21}$$

Inserendo infine la (7.21) nella (7.19) si ottiene la (7.17). □

Il disegno campionario stratificato in cui le numerosità dei campioni dei diversi strati sono quelle previste dalla (7.17) è detto *disegno stratificato con*

allocazione di Neyman, o, in breve, *campione di Neyman*. Se si fa riferimento alle quantità $a_g = n_g/n$ introdotte nella Sezione 7.3, vale la relazione: $a_g = w_g\, S_{yg} / \sum_{h=1}^{M} w_h\, S_{yh}$, $g = 1, \ldots, M$.

Fissata la numerosità campionaria totale n, il campione di Neyman seleziona da ciascuno strato un numero di unità tanto più elevato quanto più:

– è elevato w_g, ossia è alto il numero di unità che compongono lo strato;
– è elevato S_{yg}, ovvero è alta la variabilità dello strato.

Questo fatto è del tutto intuitivo, in quanto gli strati di cui è più difficile stimare la media, e che quindi richiedono un più elevato numero di unità campionarie, sono quelli di maggior variabilità, con unità più eterogenee. Quanto più elevato è il grado di eterogeneità dello strato, tanto maggiore è il numero di unità da selezionare.

Se le varianze di strato sono tutte uguali si ha $S_{y1} = \cdots = S_{yM}$, per cui il campione di Neyman si riduce a quello proporzionale. Quanto maggiore è la diversità tra le varianze degli strati, tanto più il campione di Neyman differirà da quello proporzionale.

Se si usa il campione di Neyman, la varianza dello stimatore $\widehat{\mu}_{str}$ assume il valore

$$\begin{aligned} V\left(\widehat{\mu}_{str};\ Ney\right) &= \sum_{g=1}^{M} \left\{ w_g^2 S_{yg}^2 \Big/ n \frac{w_g\, S_{yg}}{\sum_{h=1}^{M} w_h\, S_{yh}} \right\} - \frac{1}{N} \sum_{g=1}^{M} w_g\, S_{yg}^2 \\ &= \frac{1}{n} \left(\sum_{g=1}^{M} w_g\, S_{yg} \right)^2 - \frac{1}{N} \sum_{g=1}^{M} w_g\, S_{yg}^2. \end{aligned} \tag{7.22}$$

A parità di numerosità totale n, il campione di Neyman conferisce allo stimatore $\widehat{\mu}_{str}$ un'efficienza maggiore rispetto al disegno proporzionale. In altre parole, vale la relazione:

$$V(\widehat{\mu}_{str};\ Ney) \leqslant V(\widehat{\mu}_{str};\ prop).$$

Inoltre, l'uso del campione di Neyman porta vantaggi di efficienza tanto maggiori rispetto al proporzionale quanto più i termini $S_{y1}, \ldots, S_{yM}$ sono differenti tra loro.

Esempio 7.5. Consideriamo il *file* `cultura.txt`, già visto nel Capitolo 3, in cui sono riportate, per una popolazione di 1500 famiglie, l'ampiezza del nucleo familiare, il titolo di studio del capofamiglia, il reddito annuo disponibile (in Euro), le spese annue (in Euro) per attività culturali (teatro, cinema, libri e riviste, visite a muse, mostre, etc.).

Per semplificare la trattazione supponiamo di conoscere *a priori*, per ciascuna famiglia, solo il titolo di studio del capofamiglia. Supponiamo inoltre che l'interesse verta sulla stima della spesa media annua per attività culturali. La media è $\mu_y = 702.5$, e la deviazione *standard* (corretta) $S_y = 592.6$.

Tabella 7.1 Caratteristiche degli strati di una popolazione di 1500 unità

Strato g	*Numerosità* N_g	*Peso* w_g	*Media* μ_{yg}	*Deviazione standard* S_{yg}	*Prodotto* $w_g\, S_{yg}$
1 – Media inferiore	835	0.56	413.6	317.8	177.97
2 – Media superiore	395	0.26	814.0	548.1	142.51
3 – Laurea	270	0.18	1431.6	625.4	112.57
	1500	1			433.05

Tabella 7.2 Pesi campionari di strato ($a_g = n_g/n$)

Strato	*Allocazione proporzionale*	*Allocazione di Neyman*	*Allocazione uniforme*
1	0.56	0.41	$0.\overline{3}$
2	0.26	0.33	$0.\overline{3}$
3	0.18	0.26	$0.\overline{3}$

È logico ritenere che le spese per attività culturali siano legate al titolo di studio del capofamiglia. Al crescere di questo, presumibilmente si spenderà di più in attività culturali. Per questa ragione formeremo gli strati sulla base del titolo di studio del capofamiglia, che quindi viene usato come *carattere di stratificazione* . Ogni strato sarà formato da tutte le famiglie con lo stesso titolo di studio del capofamiglia. Si hanno in totale tre strati, di cui numerosità, pesi, medie e deviazioni *standard* (corrette) sono riportate in Tabella 7.1.

Come era lecito attendersi, le medie di strato crescono al crescere del titolo di studio. È interessante osservare che anche le deviazioni *standard* degli strati crescono al crescere del titolo di studio. Una situazione di questo tipo si riscontra assai di frequente nella pratica applicativa: gli strati con i valori più grandi del carattere di interesse sono quelli di più alta variabilità.

In Tabella 7.2 sono invece riportate le quantità $a_1 = n_1/n$, $a_2 = n_2/n$, $a_3 = n_3/n$, nei tre casi di allocazione proporzionale, di Neyman e uniforme (da ogni strato si estrae lo stesso numero di unità). Come si vede, l'allocazione di Neyman è piuttosto diversa da quella proporzionale, a causa delle differenti varianze degli strati.

Le quantità $\sum_g w_g S_{yg}^2$ e $\left(\sum_g w_g\, S_{yg}\right)^2$ sono rispettivamente pari a 205086.3 e 187532.3. Per valutare il guadagno di efficienza che si ottiene mediante la stratificazione, abbiamo calcolato le varianze di $\overline{\mu}_{str}$ con le tre allocazioni proporzionale, di Neyman e uniforme, per diverse numerosità campionarie totali n. Queste sono poi confrontate con la varianza della media campionaria $\overline{y}_{\mathbf{s}}$ usata in coppia con il disegno ssr. I risultati sono riportati in Tabella 7.3.

In questo caso l'uso della stratificazione porta notevoli vantaggi rispetto a quanto si ottiene con il disegno ssr. Infatti, lo stimatore $\widehat{\mu}_{str}$ usato con il disegno stratificato proporzionale ha, a parità di ampiezza del campione, una

Tabella 7.3 Varianze per disegni ssr e stratificato (allocazione proporzionale, uniforme, di Neyman)

n	$V(\overline{y}_{\mathbf{s}};\ ssr)$ (1)	$V(\widehat{\mu}_{str};\ prop)$ (2)	$V(\widehat{\mu}_{str};\ unif)$ (3)	$V(\widehat{\mu}_{str};\ Ney)$ (4)
50	6789.4	3965.0	3743.1	3613.9
75	4448.2	2597.8	2449.8	2363.7
100	3277.6	1914.1	1803.2	1738.6
150	2107.0	1230.5	1156.5	1113.5
200	1521.8	888.7	833.2	800.9
300	936.5	546.9	509.9	488.4
n		$\frac{(1)-(2)}{(1)} \times 100$	$\frac{(1)-(3)}{(1)} \times 100$	$\frac{(1)-(4)}{(1)} \times 100$
50		41.6%	44.9%	46.8%
75		41.6%	44.9%	46.9%
100		41.6%	45.0%	47.0%
150		41.6%	45.1%	47.2%
200		41.6%	45.2%	47.4%
300		41.6%	45.6%	47.9%

varianza del 41.6% più piccola di quella della media campionaria $\overline{y}_{\mathbf{s}}$ usata con il disegno ssr (si ricordi che in questo caso è $\widehat{\mu}_{str} = \overline{y}_{\mathbf{s}}$). Ulteriori vantaggi di efficienza (tutto sommato abbastanza contenuti) si ottengono usando il campione di Neyman. Si osservi che, contrariamente a quel che accade nell'Esempio 7.3, la regola di allocazione uniforme fornisce risultati migliori di quella proporzionale. Questo dipende dal fatto che nel presente esempio l'allocazione uniforme è più "simile" a quella di Neyman di quanto lo sia l'allocazione proporzionale. □

Nel problema di ottimo che genera la formula di Neyman (7.17) non è stato posto il vincolo che le numerosità campionarie degli strati non superino le corrispondenti numerosità degli strati stessi. In altre parole, non sono stati posti i vincoli $n_g \leqslant N_g$ per tutti gli strati $g = 1, \ldots, M$. Ciò implica che per qualche strato la formula (7.17) può produrre un valore di n_g maggiore di N_g, il che è assurdo. Un esame attento della (7.17) suggerisce che il problema sorge in pratica quando le varianze degli strati sono molto diverse tra loro, e la numerosità campionaria totale n è "grande".

Per ovviare al problema segnalato occorre modificare la formula di Neyman. In primo luogo, iniziamo con l'osservare che il primo strato in cui, usando la (7.17), si ha $n_g > N_g$, è null'altro che il primo strato in cui il rapporto n_g/N_g è maggiore di 1. Essendo

$$\frac{n_g}{N_g} = \frac{n}{N_g} \frac{w_g\, S_{yg}}{\sum_{h=1}^{M} w_h\, S_{yh}} = \frac{n}{N \sum_{h=1}^{M} w_h\, S_{yh}}\, S_{yg},\ g = 1 \ldots,\ M$$

il più alto valore di n_g/N_g corrisponde allo strato con il più grande valore di S_{yg}, ossia allo strato con varianza più grande. Per semplicità di trattazione assumeremo d'ora in poi di ordinare gli strati da quello con varianza più grande a quello con varianza più piccola, così che $S_{y1} \geqslant S_{y2} \geqslant \cdots \geqslant S_{yM}$. Questo significa che il primo strato in cui, in base alla (7.17), si ha $n_g > N_g$ è lo strato 1, il secondo è lo strato 2, e così via.

Non appena si ha $n_1 > N_1$ si campionano tutte le unità dello strato 1, e si ripartisce la residua numerosità campionaria totale $n - N_1$ tra i restanti strati $2, \ldots, M$ con la regola di Neyman ristretta agli strati stessi. In altre parole, si considerano numerosità campionarie di strato del tipo:

$$n_{1,1} = N_1; \;\; n_{g,1} = (n - N_1)\, \frac{w_g\, S_{yg}}{\sum_{h=2}^{M} w_h\, S_{yh}}, \;\; g = 2, \ldots, M.$$

Questa regola di allocazione vale ovviamente finché $n_{g,1} \leqslant N_g$, $g = 2, \ldots, M$. Non appena risulta $n_{2,1} > N_2$, le numerosità campionarie di strato saranno del tipo:

$$n_{1,2} = N_1; \;\; n_{2,2} = N_2; \;\; n_{g,2} = (n - N_1 - N_2)\, \frac{w_g\, S_{yg}}{\sum_{h=3}^{M} w_h\, S_{yh}}, \;\; g = 3, \ldots, M,$$

e così via. Questo tipo di aggiustamento viene effettuato fino a quando le numerosità campionarie di strato risultano tutte inferiori (o uguali) alle corrispondenti ampiezze degli strati stessi.

Esempio 7.6. Si consideri una popolazione finita di $N = 2000$ unità, suddivise in tre strati di numerosità rispettivamente $N_1 = 400$, $N_2 = 1000$, $N_3 = 600$. Si supponga anche che le deviazioni standard (corrette) degli strati siano $S_{y1} = 1200$, $S_{y2} = 500$, $S_{y3} = 100$. Gli strati sono già ordinati da quello con varianza massima a quello con varianza minima, per cui non c'è bisogno di riordinarli e rinumerarli.

Poiché i pesi degli strati sono $w_1 = 0.2$, $w_2 = 0.5$, $w_3 = 0.3$, si ha $w_1\, S_{y1} = 240$, $w_2\, S_{y2} = 250$, $w_3\, S_{y3} = 30$, da cui $\sum_{g=1}^{3} w_g\, S_{yg} = 520$. Se la numerosità campionaria è n, in base alla (7.17) le numerosità campionarie di strato devono essere del tipo:

$$n_1 = \frac{240}{520}\, n = 0.462\, n, \;\; n_2 = \frac{250}{520}\, n = 0.480\, n, \;\; n_3 = \frac{30}{520}\, n = 0.058\, n. \quad (7.23)$$

Le (7.23) non valgono per ogni numerosità campionaria n, ma solo per valori di n tali che $n_g \leqslant N_g$. Come detto, il primo strato in cui è $n_g > N_g$ è quello con varianza più grande, ossia il primo. In effetti, si ha $n_1 \leqslant 400$ se e solo se $0.462\, n \leqslant 400$, cioè se e solo se $n \leqslant 866$. Per $n \geqslant 867$ si ha invece $0.462\, n \geqslant 401$. Applicando i ragionamenti sopra svolti, per $n \geqslant 867$ l'allocazione di Neyman effettiva sarà del tipo:

$$
\begin{aligned}
n_{1,1} &= 400 \\
n_{2,1} &= \frac{250}{250+30}\,(n-400) = 0.893\,(n-400) \\
n_{3,1} &= \frac{30}{250+30}\,(n-400) = 0.107\,(n-400).
\end{aligned}
\tag{7.24}
$$

Le (7.24) non vale per tutte le $n \geqslant 867$, ma solo fino a quando $n_{2,1} \leqslant N_2$, ovvero fino a quando $0.893\,(n-400) \leqslant 1000$. Questo significa che le (7.24) valgono solo per $867 \leqslant n \leqslant 1520$. Per $n \geqslant 1521$ l'allocazione di Neyman effettiva sarà invece

$$
\begin{aligned}
n_{1,2} &= 400 \\
n_{2,2} &= 1000 \\
n_{3,1} &= \frac{30}{30}\,(n-400-1000) = n-1400.
\end{aligned}
$$

□

7.4.2 Allocazione ottima per una data funzione di costo

Le considerazioni che hanno portato al campione di Neyman possono anche essere usate nel caso in cui non sia fissato il numero totale di unità campionarie, ma piuttosto l'ammontare massimo spendibile per effettuare la rilevazione statistica. Supponiamo che l'ammontare di denaro a disposizione per effettuare la rilevazione (mediante un disegno stratificato) sia pari a C, e che il costo di rilevazione si possa dividere in due parti:

- un costo fisso c_0;
- un costo variabile dipendente dal numero di unità che si campionano da ogni strato.

Per quanto riguarda la parte variabile del costo di rilevazione (che ovviamente è la più importante) si può assumere che il costo di campionamento e osservazione di un'unità dello strato g-mo sia pari a c_g, così che per selezionare n_g unità si deve sostenere un costo pari a $c_g\,n_g$. Il costo totale di rilevazione è quindi eguale a:

$$
c_0 + \sum_{g=1}^{M} c_g\,n_g.
$$

Un criterio molto naturale per stabilire le numerosità campionarie degli strati è quello di determinare $n_1, \ldots, n_M$ in modo da rendere minima la varianza $V(\widehat{\mu}_{str})$, con il vincolo che il costo di rilevazione sia C. Formalmente, si ha il seguente problema di minimo vincolato

$$
\begin{cases}
\textit{minimizzare}: \ \sum_{g=1}^{M} \frac{w_g^2 S_{yg}^2}{n_g} \\
\textit{con il vincolo}: \ c_0 + \sum_{g=1}^{M} c_g\,n_g = C
\end{cases}.
\tag{7.25}
$$

Proposizione 7.4. *La soluzione del problema di ottimo* (7.25) *assume la forma*

$$n_g = (C - c_0)\,\frac{w_g\,S_{yg}/\sqrt{c_g}}{\sum_{h=1}^{M}\sqrt{c_h}\,w_h\,S_{yh}};\ g = 1,\,\ldots,\,M. \tag{7.26}$$

Dimostrazione. La determinazione della soluzione del problema (7.25) è del tutto simile a quella del problema (7.16). In primo luogo, la funzione Lagrangiana è del tipo

$$\mathcal{L}(n_1,\,\ldots,\,n_M\ \lambda) = \sum_{g=1}^{M}\frac{w_g^2 S_{yg}^2}{n_g} + \lambda\,\left(c_0 + \sum_{g=1}^{M} c_g\,n_g - C\right).$$

Calcolando le sue derivate rispetto a $n_1,\,\ldots,\,n_M,\,\lambda$ e annullando tali derivate, si ottengono le equazioni

$$\frac{\partial\mathcal{L}}{\partial n_g} = -\frac{w_g^2 S_{yg}^2}{n_g^2} + \lambda c_g = 0\ \ da\ cui$$

$$\sqrt{c_g}\,S_{yg}\,w_g = \sqrt{\lambda}\,c_g\,n_g;\ g = 1,\,\ldots,\,M \tag{7.27}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = c_0 + \sum_{g=1}^{M} c_g\,n_g - C = 0\ \ da\ cui\ \ \sum_{g=1}^{M} c_g\,n_g = C - c_0. \tag{7.28}$$

Sommando membro a membro le (7.27) e sfruttando la (7.28) si ha poi la relazione

$$\sum_{g=1}^{M}\sqrt{c_g}\,w_g\,S_{yg} = \sqrt{\lambda}\sum_{g=1}^{M} c_g n_g = \sqrt{\lambda}(C - c_0)$$

da cui si ottiene

$$\sqrt{\lambda} = \frac{1}{C - c_0}\sum_{g=1}^{M}\sqrt{c_g}\,w_g\,S_{yg}. \tag{7.29}$$

Dalle (7.27) e (7.29) è a questo punto facile ottenere la (7.26). □

Se le numerosità campionarie di strato sono scelte in base alla (7.26), la dimensione totale del campione sarà pari a

$$n = \sum_{g=1}^{M} n_g = (C - c_0)\,\frac{\sum_{g=1}^{M} w_g\,S_{yg}/\sqrt{c_g}}{\sum_{g=1}^{M}\sqrt{c_g}\,w_g\,S_{yg}}.$$

Pertanto, se si fa riferimento riferimento alle quantità $a_g = n_g/n$ si ha in questo caso: $a_g = w_g\,S_{yg}/\sqrt{c_g}/\sum_{h=1}^{M} w_h\,S_{yh}/\sqrt{c_h},\ g = 1,\,\ldots,\,M.$

È facile verificare (Esercizio 7.2) che se i costi di selezione e osservazione dai diversi strati sono tutti uguali ($c_1 = \cdots = c_M$), la (7.26) si riduce all'usuale allocazione di Neyman.

Per determinare il campione ottimo in base alla nostra funzione di costo non sono posti i vincoli $n_g \leqslant N_g$, $g = 1, \ldots, M$. Questo significa che è necessario aggiustare la formula (7.26) similmente a quanto fatto per il campione di Neyman.

7.4.3 Considerazioni sul caso in cui le varianze degli strati siano incognite

Per poter utilizzare l'allocazione di Neyman (7.17) (ma gli stessi argomenti valgono per il campione ottimo di costo dato) è necessario conoscere le quantità $S_{y1}, \ldots, S_{yM}$, ossia, in sostanza, le varianze degli strati. Questo è un punto assai delicato in quanto il caso più frequente è quello in cui le varianze di strato sono incognite.

In assenza di informazioni extra-campionarie sulle varianze di strato, si potrebbe pensare di usare un metodo a due fasi per ottenere una stima preliminare di $S_{y1}, \ldots, S_{yM}$. Precisamente, data la numerosità totale n si può procedere nel modo seguente.

- *Fase* 1. Scelta una numerosità campionaria iniziale $n_p < n$, si seleziona dalla popolazione un campione stratificato proporzionale, con numerosità campionarie di strato del tipo $n_{p1} = n_p\, w_1, \ldots, n_{pM} = n_p\, w_M$. Sulla base di tale campione si possono ottenere (secondo quanto esposto nella Sezione 7.2) delle stime $\widehat{s}^2_{p\,y1}, \ldots, \widehat{s}^2_{p\,yM}$ rispettivamente di $S^2_{y1}, \ldots, S^2_{yM}$.
- *Fase* 2. Si calcolano le numerosità campionarie di Neyman in cui però al posto delle incognite $S_{y1}, \ldots, S_{yM}$ si usano le loro stime $\widehat{s}_{p\,y1}, \ldots, \widehat{s}_{p\,yM}$. In altre parole, si calcolano le quantità

$$n_g = n \frac{w_g\, \widehat{s}_{p\,yg}}{\sum_{h=1}^{M} w_h\, \widehat{s}_{p\,yh}};\ g = 1, \ldots, M \tag{7.30}$$

  e si usano queste come numerosità campionarie di strato. Poiché dal generico strato g sono già selezionate n_{pg} unità ($g = 1, \ldots, M$), verranno da esso selezionate ulteriori $n_g - n_{pg}$ unità "residue", in modo che le ampiezze campionarie degli strati siano quelle previste dalla (7.30).

Per quanto riguarda la numerosità campionaria n_p, essa dovrebbe essere abbastanza piccola, ma comunque tale da fornire stime abbastanza accurate delle varianze di strato. In genere la scelta di n_p avviene sulla base di considerazioni di costo. Molto comune è il caso in cui si prende n_p pari al 10% o al 20% dell'ampiezza campionaria totale n.

Usando la procedura a due fasi sopra descritta la formula della varianza (7.22) non è più valida, sia perché non si usano le "vere" varianze di strato, sia perché le numerosità campionarie (7.30) dipendono dal campione selezionato

nella prima fase, e quindi dagli stessi dati campionari. Quest'ultimo fatto implica che neanche la (7.7) rappresenta, a stretto rigore, la varianza di $\widehat{\mu}_{str}$. Essa si può usare solo a titolo di approssimazione della vera $V(\widehat{\mu}_{str})$. Per le stesse ragioni, la (7.9) non è più uno stimatore corretto di $V(\widehat{\mu}_{str})$. Ad ogni modo, si continuerà anche in questo caso a stimare $V(\widehat{\mu}_{str})$ con la (7.9).

L'utilizzo della procedura a due fasi del tipo sopra descritto ha effettivamente senso soltanto quando ci si aspetta che l'allocazione ottimale produca un netto miglioramento di efficienza rispetto all'allocazione proporzionale. Ora, vi sono due casi in cui questo di sicuro accade. Il primo è quello in cui i costi di osservazione delle unità sono molto diversi da uno strato all'altro, ed è fissata la spesa totale C da sostenere per la rilevazione campionaria. In questo caso i valori delle c_g possono rendere le numerosità campionarie che si ottengono dalla regola di allocazione ottimale della Sezione 7.4.2 molto diverse da quelle basate sull'allocazione proporzionale, con una conseguente sensibile diminuzione della varianza di $\widehat{\mu}_{str}$. Il secondo caso è quello in cui è noto che le varianze di strato sono molto diverse tra loro. L'uso dell'allocazione ottimale (campione di Neyman) può risultare molto vantaggioso rispetto all'allocazione proporzionale. Naturalmente, queste considerazioni vanno fatte tenendo ben presente che non si usano i veri valori delle S_{yg}, ma solo delle loro stime da campione pilota, che da esse possono differire non poco. Il ricorso a valori non adeguati per le varianze di strato può condurre ad una perdita di precisione delle stime rispetto al campionamento casuale semplice.

Esempio 7.7. Consideriamo ancora la popolazione dell'Esempio 7.5, e supponiamo di voler selezionare un campione stratificato di $n = 250$ unità. Per cercare di approssimare l'allocazione di Neyman si può procedere con il meccanismo a due fasi sopra descritto.

Nella prima fase selezioniamo un campione pilota, di tipo stratificato proporzionale, di numerosità $n_p = 50$. Le numerosità campionarie di strato del campione pilota sono indicate qui di seguito:

$$n_{p1} = 0.56 \times 50 = 28,\ n_{p2} = 0.26 \times 50 = 13,\ n_{p3} = 0.18 \times 50 = 9.$$

I dati campionari ottenuti sono contenuti nei *file* `cp1.txt`, `cp2.txt`, `cp3.txt`. Come stime delle deviazioni *standard* di strato abbiamo le seguenti:

$$\widehat{s}_{p\,y1} = 278.9,\ \widehat{s}_{p\,y2} = 542.4,\ \widehat{s}_{p\,y3} = 488.5.$$

Sulla base di questi risultati, nella seconda fase calcoliamo le numerosità campionarie di strato con la formula di Neyman, in cui le S_{yg} sono rimpiazzate dalle $\widehat{s}_{p\,yg}$. Si ha:

$$n_1 = \frac{0.56 \times 278.9}{0.56 \times 278.9 + 0.26 \times 542.4 + 0.18 \times 488.5} \times 250 = 101,$$
$$n_2 = \frac{0.26 \times 542.4}{0.56 \times 278.9 + 0.26 \times 542.4 + 0.18 \times 488.5} \times 250 = 92,$$
$$n_3 = \frac{0.18 \times 488.5}{0.56 \times 278.9 + 0.26 \times 542.4 + 0.18 \times 488.5} \times 250 = 57.$$

Rimangono da estrarre 73 unità "residue" dallo strato 1, 79 unità "residue" dallo strato 2, e 48 unità "residue" dallo strato 3. I *file* `cr1.txt`, `cr2.txt`, `cr3.txt` contengono i dati campionari residui, e i *file* `ct1.txt`, `ct2.txt`, `ct3.txt` contengono i dati campionari totali, rispettivamente per lo strato 1, lo strato 2, e lo strato 3.

Le medie e le varianze campionarie di strato, relativamente ai campioni "totali", sono riportate qui di seguito:

$$\overline{y}_1 = 390.2,\ \overline{y}_2 = 839.0,\ \overline{y}_3 = 1333.4;\ \widehat{s}^2_{y1} = 93271.1,\ \widehat{s}^2_{y2} = 290508.2,\ \widehat{s}^2_{y3} = 450855.0\,.$$

Da esse si ottiene la seguente stima della media della popolazione:

$$\begin{aligned}\widehat{\mu}_{str} &= w_1\,\overline{y}_1 + w_2\,\overline{y}_2 + w_3\,\overline{y}_3\\ &= 0.56 \times 390.2 + 0.26 \times 839.0 + 0.18 \times 1333.4\\ &= 676.7\,.\end{aligned}$$

Per stimare $V(\widehat{\mu}_{str})$ usiamo infine la (7.9), che è pari a

$$\begin{aligned}\widehat{V}(\widehat{\mu}_{str}) &= \left(\frac{1}{n_1} - \frac{1}{N_1}\right) w_1^2 \widehat{s}^2_{y1} + \left(\frac{1}{n_2} - \frac{1}{N_2}\right) w_2^2 \widehat{s}^2_{y2} + \left(\frac{1}{n_3} - \frac{1}{N_3}\right) w_3^2 \widehat{s}^2_{y3}\\ &= \left(\frac{1}{101} - \frac{1}{835}\right) \times 0.56^2 \times 93271.1 + \left(\frac{1}{92} - \frac{1}{395}\right) \times 0.26^2 \times 290508.2\\ &\quad + \left(\frac{1}{57} - \frac{1}{270}\right) \times 0.18^2 \times 450855.0\\ &= 620.5\,.\end{aligned}$$

□

7.5 Scelta della numerosità campionaria

Nella sezione precedente si è ampiamente trattato del problema dell'allocazione delle unità campionarie ai vari strati, data la numerosità campionaria totale n. Come risulta chiaro dalla trattazione precedente, questo equivale a scegliere le quantità $a_g = n_g/n$, $g = 1, \ldots, M$, dato n. Nella presente sezione verrà brevemente trattato il problema della scelta di n.

Se è fissato l'ammontare totale C di denaro disponibile per effettuare la rilevazione, la numerosità campionaria sarà da esso determinata. Infatti, dati i valori di $a_1, \ldots, a_M$, e ricordando che $n_g = n\,a_g$, il costo di rilevazione è pari a

$$c_0 + \sum_{g=1}^{M} c_g\,n_g = c_0 + n \sum_{g=1}^{M} c_g\,a_g.$$

Poiché tale costo deve essere uguale a C, deve valere la relazione $c_0 + n \sum_g c_g\, a_g = C$, dalla quale si desume che

$$n = \frac{C - c_0}{\sum_{g=1}^{M} c_g\, a_g}.$$

Se invece la struttura dei costi di rilevazione non è chiaramente specificata, oppure se non è fissato un tetto massimo di spesa C, bisogna percorrere altre strade. La più semplice è quella di procedere sulla falsariga di quanto già visto per il disegno ssr.

In primo luogo, dati $a_1, \ldots, a_M$ la varianza dello stimatore $\widehat{\mu}_{str}$ si può scrivere come

$$\begin{aligned} V(\widehat{\mu}_{str}) &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n\, a_g} - \frac{1}{N\, w_g} \right) S_{yg}^2 \\ &= \frac{1}{n} \sum_{g=1}^{M} \frac{w_g^2 S_{yg}^2}{a_g} - \frac{1}{N} \sum_{g=1}^{M} w_g\, S_{yg}^2 = \frac{1}{n}\, V - \frac{1}{N}\, V_0 \end{aligned} \tag{7.31}$$

dove, per comodità di notazione, si è posto $V = \sum_{g=1}^{M} \frac{w_g^2 S_{yg}^2}{a_g}$, e $V_0 = \sum_{g=1}^{M} w_g\, S_{yg}^2$.

L'obiettivo è quello di determinare la numerosità campionaria totale n in maniera tale che l'errore assoluto di stima $|\widehat{\mu}_{str} - \mu_y|$ sia superiore ad una soglia t con probabilità pari a α, con t, α fissati a priori. In simboli:

$$Pr\left(|\widehat{\mu}_{str} - \mu_y| > t\right) = \alpha. \tag{7.32}$$

Per quanto riguarda la distribuzione di probabilità di $\widehat{\mu}_{str}$, essa verrà approssimata con una normale di media μ_y e varianza (7.31). Questo significa che la v.a. standardizzata

$$\frac{\widehat{\mu}_{str} - \mu_y}{\sqrt{\frac{1}{n}\, V - \frac{1}{N}\, V_0}}$$

ha in via approssimata distribuzione normale standard $N(0,\, 1)$. La (7.32) si può allora riscrivere come

$$\begin{aligned} Pr\left(|\widehat{\mu}_{str} - \mu_y| > t\right) &= Pr\left(\frac{|\widehat{\mu}_{str} - \mu_y|}{\sqrt{\frac{1}{n}\, V - \frac{1}{N}\, V_0}} > \frac{t}{\sqrt{\frac{1}{n}\, V - \frac{1}{N}\, V_0}} \right) \\ &\approx Pr\left(|N(0,\, 1)| > \frac{t}{\sqrt{\frac{1}{n}\, V - \frac{1}{N}\, V_0}} \right) \\ &= 2Pr\left(N(0,\, 1) > \frac{t}{\sqrt{\frac{1}{n}\, V - \frac{1}{N}\, V_0}} \right) \\ &= \alpha \end{aligned}$$

da cui si trae la relazione

$$Pr\left(N(0,\,1) > \frac{t}{\sqrt{\frac{1}{n}\,V - \frac{1}{N}\,V_0}}\right) = \frac{\alpha}{2}\,. \tag{7.33}$$

Usando i soliti ragionamenti (e i soliti simboli), dalla (7.33) si desume che

$$\frac{t}{\sqrt{\frac{1}{n}\,V - \frac{1}{N}\,V_0}} = z_{\alpha/2}$$

da cui, con facili passaggi, si ottiene la seguente espressione per la numerosità campionaria:

$$n = \frac{\frac{z^2_{\alpha/2}}{t^2}\,V}{1 + \frac{1}{N}\frac{z^2_{\alpha/2}}{t^2}\,V_0}. \tag{7.34}$$

Per N "grande" il termine $(z_{\alpha/2}/t)^2 V_0/N$ è pressoché trascurabile, per cui la (7.34) si riduce a $n = (z_{\alpha/2}/t)^2\,V$.

Per poter effettivamente usare la (7.34) è necessario disporre di informazioni sulle S^2_{yg}, ossia, in buona sostanza, sulle varianze degli strati. In assenza di tali informazioni si può utilizzare la tecnica del campione pilota descritto nella Sezione 7.4.3. Al termine della prima fase, sulla base delle stime $\widehat{s}^2_{p\,y1}, \ldots, \widehat{s}^2_{p\,yM}$ si costruiscono stime di V e V_0, le quali vengono poi usate nella (7.34). Naturalmente, la seconda fase del metodo descritto nella Sezione 7.4.3 rimane invariata.

Esempio 7.8. Consideriamo ancora la popolazione di 1500 famiglie dell'Esempio 7.5, e supponiamo di conoscere le varianze (corrette) degli strati. L'obiettivo è quello di determinare la numerosità campionaria n in modo che l'errore assoluto di stima $|\widehat{\mu}_{str} - \mu_y|$ sia superiore a 40 Euro con probabilità 0.05. Questo significa, con i simboli in precedenza adottati, che $t = 40$, $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$. Dall'Esempio 7.5 già sappiamo che $V = \sum_g w_g\,S^2_{yg} = 205068.3$. Se si adotta l'allocazione di Neyman, essendo $a_g = w_g\,S_{yg}/\sum_h w_h\,S_{yh}$, si ha $V = \left(\sum_g w_g\,S_{yg}\right)^2 = 187532.3$. Dalla (7.34) si ricava pertanto una numerosità campionaria pari a:

$$n = \frac{\frac{1.96^2}{40^2} \times 187532.3}{1 + \frac{1}{1500} \times \frac{1.96^2}{40^2} \times 205068.3} = 339\,. \qquad \square$$

Esempio 7.9. Se nell'Esempio precedente non si hanno informazioni sulle varianze di strato, per scegliere n si può utilizzare la tecnica del campione pilota. Nel presente caso usiamo i risultati già descritti nell'Esempio 7.7, relativi ad

un campione pilota con $n_p = 50$. La regola di allocazione usata è ancora quella di Neyman, ma con le "vere" deviazioni *standard* S_{yg} sostituite dalle loro stime $\widehat{s}_{p\,yg}$ ottenute grazie al campione pilota. Le due quantità V_0, V possono essere stimate nel modo seguente

$$\begin{aligned}\widehat{V}_{p0} &= \sum_{g=1}^{3} w_g\, \widehat{s}^{\,2}_{p\,yg} \\ &= 163002.2, \\ \widehat{V}_{p} &= \left(\sum_{g=1}^{3} w_g\, \widehat{s}_{p\,yg}\right)^2 \\ &= 148331.3\,.\end{aligned}$$

Dalla (7.34) si ricava pertanto una numerosità campionaria richiesta pari a:

$$n = \frac{\frac{1.96^2}{40^2} \times 148331.3}{1 + \frac{1}{1500} \times \frac{1.96^2}{40^2} \times 163002.2} = 339\,. \qquad \square$$

Come alternativa al metodo di scelta di n dianzi descritto, ci si può basare sull'effetto del disegno, introdotto nel Capitolo 3. Dati $a_1, \ldots, a_M$, l'effetto del disegno è pari a:

$$\mathit{Deff}\,(str,\ \widehat{\mu}_{str}) = \frac{V(\widehat{\mu}_{str};\ str)}{V(\overline{y}_{\mathbf{s}};\ ssr)} = \frac{\frac{1}{n}\,V - \frac{1}{N}\,V_0}{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}.$$

Se, sulla base di precedenti rilevazioni o di un campione pilota, è noto a priori il valore di $\mathit{Deff}\,(str,\ \widehat{\mu}_{str})$ (o almeno una sua stima sufficientemente accurata) ci si può basare su di esso per scegliere la numerosità campionaria n. Il procedimento è molto semplice, e consta di due fasi:

- fissati i valori di t e di α, si determina la numerosità campionaria n_{ssr} necessaria affinché, con un disegno di tipo semplice senza ripetizioni, sia $Pr(|\overline{y}_{\mathbf{s}} - \mu_y| > t) = \alpha$, secondo le linee esposte nel Capitolo 4;
- si calcola $n = n_{ssr}\,\mathit{Deff}\,(str,\ \widehat{\mu}_{str})$, che fornisce la numerosità campionaria richiesta.

7.6 Alcuni principî di base per la costruzione di strati

L'obiettivo di base della stratificazione di una popolazione, come più volte rimarcato, è quello di formare gruppi di unità il più possibile *omogenei* dal punto di vista del carattere $\mathcal{Y}$ oggetto di interesse. Gli strati, in altre parole, dovrebbero essere formati da unità con modalità y "simili" . Questo significa che ci dovrebbe essere poca variabilità dentro gli strati, e che il grosso della variabilità del carattere $\mathcal{Y}$ dovrebbe essere *tra* gli strati, ossia tra le medie

degli strati stessi. Per raggiungere un simile obiettivo la situazione ideale è quella in cui le modalità y sono note per le unità della popolazione. In tal caso, il primo strato potrebbe essere formato dalle unità con le modalità y più piccole, il secondo strato dalle unità con modalità y intermedie, e così via. Formalmente, questo significa stabilire dei "punti di taglio" (*cutpoint*) $b_1, \ldots, b_{M-1}$, e definire gli strati nel modo seguente:

- strato 1: insieme delle unità i con modalità $y_i \leqslant b_1$;
- strato 2: insieme delle unità i con modalità $b_1 < y_i \leqslant b_2$;
  ...
- strato $M-1$: insieme delle unità i con modalità $b_{M-2} < y_i \leqslant b_{M-1}$;
- strato M: insieme delle unità i con modalità $y_i > b_{M-1}$.

Il problema è (ovvio!) che le modalità del carattere $\mathcal{Y}$ sono incognite. Dopotutto, se le conoscessimo non avremmo alcun bisogno di ricorrere ad una rilevazione campionaria. Per questa ragione, la costruzione di strati è in genere effettuata tramite caratteri statistici ausiliari, in genere denominati, visto il contesto, *caratteri di stratificazione* . L'idea di base è molto semplice: a valori simili dei caratteri di stratificazione dovrebbero corrispondere valori simili del carattere di interesse $\mathcal{Y}$. In tal modo, la formazione degli strati sulla base delle variabili di stratificazione dovrebbe produrre strati abbastanza omogenei dal punto di vista della variabile di interesse $\mathcal{Y}$. Affinché questo modo di procedere sia efficace, è necessario che vi sia uno stretto legame tra i caratteri di stratificazione e quello di interesse. Nel caso di legami deboli o assenti, la stratificazione della popolazione sarebbe del tutto inutile, in quanto produrrebbe strati di forte disomogeneità e quindi scarsamente utili. Questo è un punto assai importante in quanto, in genere, la suddivisione di una popolazione in strati richiede un lavoro assai oneroso, in termini di tempo e di costi, per reperire i caratteri di stratificazione e raggruppare le unità della popolazione stessa in base ai loro valori. Questo significa che vale la pena effettuare il lavoro di stratificazione soltanto quando si è ragionevolmente sicuri che esso produca un buon guadagno in termini di efficienza di stima.

La scelta dei caratteri di stratificazione è un problema assai rilevante, ed è ovviamente legata al tipo di rilevazione che si effettua. Spesso le unità della popolazione sono entità fisicamente esistenti in un dato territorio, che vengono stratificate sulla base di un criterio di contiguità territoriale. Questo si applica soprattutto al caso in cui le unità da campionare sono entità quali comuni o ripartizioni sub-comunali, per le quali si può ritenere che la vicinanza territoriale costituisca un fattore di omogeneità rispetto al carattere $\mathcal{Y}$ di interesse. Inoltre (e questo non è un fatto da poco) la formazione di strati di unità fisicamente "vicine" facilita sia la selezione che l'osservazione delle unità, riducendo i costi di rilevazione. L'osservazione di unità fisicamente vicine, infatti, richiede tempi più contenuti rispetto al caso di unità sparse sul territorio, e consente (in parecchi casi) un miglior utilizzo del personale addetto alla rilevazione. Altre volte le unità della popolazione vengono raggruppate sulla base di caratteri di *struttura*, che si suppone abbiano influenza

sul carattere di interesse. È del tutto ovvio che i caratteri di struttura usati dipendono dal tipo di rilevazione che si effettua, in quanto dovrebbero essere legati al carattere di interesse $\mathcal{Y}$. Per esempio, in indagini sull'atteggiamento dei consumatori nei confronti di certi prodotti le unità (individui o famiglie) vengono spesso raggruppate sulla base di caratteri quali sesso, fascia di età, numero di componenti, condizioni socio-economiche, etc..

Il numero degli strati dipende da parecchi differenti elementi. In generale, si può dire che quanto maggiore è l'informazione *a priori* di cui si dispone per stratificare la popolazione, tanto maggiore è il numero degli strati che ha senso considerare. Se l'informazione di cui si dispone è scarsa o poco affidabile, gli strati costruiti in base ad essa daranno poche garanzie in termini di omogeneità interna. Solo quando si dispone di informazioni *a priori* precise ha senso costruire un numero elevato di strati. Tuttavia, l'aumento del numero di strati se da una parte comporta un aumento del grado di omogeneità degli strati stessi, dall'altra implica una riduzione delle numerosità campionarie di strato n_g con un conseguente aumento nella variabilità delle stime.

Esempio 7.10. Nel seguito viene brevemente descritto uno studio, svolto nel 1973, sull'atteggiamento degli automobilisti di Birmingham (UK), volto in particolare (ma non esclusivamente) a valutare l'uso di cinture di sicurezza. Gli aspetti principali della rilevazione sono descritti in Golder e Yeomans (1973).

Non essendo disponibile una lista degli automobilisti (al massimo, con costi e tempi elevati, si poteva ottenere una lista di possessori di automobili), la loro selezione fu attuata con una procedura a più stadi, e secondo un criterio territoriale. La città di Birmingham è composta da 39 circoscrizioni (*ward*), a loro volta suddivise in distretti (*district polling*). Lo schema di selezione, come detto, è per stadi successivi, di seguito brevemente descritti:

- al primo stadio si seleziona un campione di circoscrizioni;
- da ciascuna delle circoscrizioni selezionate al primo stadio si estrae un campione di distretti;
- da ciascuno dei distretti selezionati al secondo stadio si seleziona un campione di famiglie;
- ciascuna delle famiglie selezionate al terzo stadio viene contattata, e si intervistano tutti gli automobilisti che la compongono.

Nel seguito ci si concentrerà sul primo stadio di selezione, quello relativo alle circoscrizioni. L'idea di base è di suddividere le circoscrizioni in strati, e di selezionare da ciascuno strato un campione ssr di circoscrizioni. Un buon guadagno di efficienza dovrebbe ottenersi costruendo strati quanto più possibile omogenei dal punto di vista del comportamento degli automobilisti. Questo significa determinare, sulla base delle informazioni *a priori* disponibili sulla popolazione, che bisogna isolare dei caratteri di stratificazione sulla base dei quali formare gruppi omogenei di circoscrizioni. I caratteri usati sono tredici, elencati in Tabella 7.4 (la classe sociale 1 indica la più alta, e la 5 la più bassa). Di essi sono anche riportati la media e la deviazione standard.

Tabella 7.4 Variabili usate per stratificare le circoscrizioni

Carattere	*Media*	*Deviazione standard*
$\mathcal{X}_1$ = Famiglie con una o più automobili (%)	39.15	11.68
$\mathcal{X}_2$ = Numero medio di automobili per 100 famiglie	44.56	14.23
$\mathcal{X}_3$ = Famiglie – proprietari di casa (%)	38.77	19.09
$\mathcal{X}_4$ = Famiglie – inquilini di case enti pubblici (%)	41.36	23.21
$\mathcal{X}_5$ = Famiglie – inquilini di case private (%)	17.26	10.73
$\mathcal{X}_6$ = Famiglie – classi sociali 1 e 2 (%)	12.44	6.91
$\mathcal{X}_7$ = Famiglie – classe sociale 3 (%)	54.92	5.50
$\mathcal{X}_8$ = Famiglie – classi sociali 4 e 5 (%)	32.49	8.66
$\mathcal{X}_9$ = Popolazione di età 15–24 anni (%)	15.85	1.58
$\mathcal{X}_{10}$ = Popolazione di età 25–44 anni (%)	24.13	2.10
$\mathcal{X}_{11}$ = Popolazione di età 45–64 anni (%)	26.23	5.47
$\mathcal{X}_{12}$ = Popolazione di età 65 anni e oltre (%)	10.87	2.50
$\mathcal{X}_{13}$ = Popolazione femminile (%)	50.62	1.63

Considerazioni basate su un'analisi statistica multivariata condotta sui dati disponibili hanno portato alla decisione di considerare cinque strati. La formazione degli strati viene effettuata aggregando le circoscrizioni con valori "simili" dei caratteri di stratificazione di Tabella 7.4. Sul piano formale si è utilizzata una procedura di *cluster analysis*. Detta x_{ji} la modalità del carattere $\mathcal{X}_j$ per l'unità i ($j = 1, \ldots, 13$, $i = 1, \ldots, N$), e detta μ_{jg} la media del carattere $\mathcal{X}_j$ nello strato g-mo ($g = 1, \ldots, 5$), si è usato un semplice algoritmo iterativo. Ad ogni iterazione si considera un'unità (circoscrizione), la quale è attribuita al gruppo g per il quale la quantità

$$\sum_j (x_{ji} - \mu_{jg})^2$$

è minima. La procedura continua fino a che non vi è nessuno spostamento di unità da un gruppo all'altro, o fino a che non sia raggiunto un prefissato numero di iterazioni. I gruppi formati in questo modo costituiscono gli strati del primo stadio di campionamento.

Come osservazione generale, la stratificazione mediante procedure di *cluster analysis* si rivela particolarmente utile quando si hanno parecchi caratteri di stratificazione, tutti in varia misura influenti sul carattere di interesse. Naturalmente, molta attenzione va posta sia sul metodo utilizzato per la *cluster analysis*, sia sull'eliminazione dell'effetto dell'unità di misura sui caratteri di stratificazione, ad es. mediante la loro preventiva standardizzazione. □

7.7 Stima della varianza della popolazione*

Finora ci siamo occupati della stima della media μ_y della popolazione. Essa, d'altra parte, non è il solo parametro di interesse in rilevazioni campionarie (anche se certamente è il più importante). Un problema di un certo rilievo è quello della stima della varianza σ_y^2 della popolazione. L'obiettivo della presente sezione è la costruzione di uno stimatore corretto di σ_y^2, quando il disegno campionario è di tipo stratificato.

In primo luogo, iniziamo con l'osservare che la varianza σ_y^2 si può scrivere nella forma:

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g}(y_{gi}-\mu_y)^2 = \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g} y_{gi}^2 - \mu_y^2 \\ &= \sum_{g=1}^{M}\frac{N_g}{N}\left(\frac{1}{N_g}\sum_{i=1}^{N_g} y_{gi}^2\right) - \mu_y^2 \\ &= \sum_{g=1}^{M} w_g\,\mu_{2g} - \mu_y^2 \end{aligned} \tag{7.35}$$

dove si è posto

$$\mu_{2g} = \frac{1}{N_g}\sum_{i=1}^{N_g} y_{gi}^2,\ g = 1, \ldots, M.$$

Da risultati noti validi per il disegno ssr discende che

$$m_{2g} = \frac{1}{n_g}\sum_{i\in \mathbf{s}_g} y_{gi}^2$$

è uno stimatore corretto di μ_{2g}, qualunque sia $g = 1 \ldots, M$. Ne consegue che

$$m_2 = \sum_{g=1}^{M} w_g\, m_{2g} \tag{7.36}$$

è uno stimatore corretto di $\sum_g w_g\,\mu_{2g}$.

Rimane da costruire uno stimatore corretto del quadrato della media della popolazione, μ_y^2. A questo proposito osserviamo che, essendo $\widehat{\mu}_{str}$ corretto, si ha

$$V(\widehat{\mu}_{str}) = E\left[\widehat{\mu}_{str}^2\right] - E\left[\widehat{\mu}_{str}\right]^2 = E\left[\widehat{\mu}_{str}^2\right] - \mu_y^2$$

da cui discende che

$$\mu_y^2 = E\left[\widehat{\mu}_{str}^2\right] - V(\widehat{\mu}_{str}). \tag{7.37}$$

Ora, uno stimatore corretto di $V(\widehat{\mu}_{str})$ è lo stimatore $\widehat{V}(\widehat{\mu}_{str})$ costruito nella Sezione 7.2, (7.9). Inoltre, uno stimatore corretto di $E\left[\widehat{\mu}_{str}^2\right]$ è di sicuro $\widehat{\mu}_{str}^2$. Sulla base della (7.37) si conclude che

$$\widehat{\mu}_y^2 = \widehat{\mu}_{str}^2 - \widehat{V}(\widehat{\mu}_{str}). \tag{7.38}$$

Usando a questo punto la (7.35) abbiamo provato la seguente proposizione.

Proposizione 7.5. *Se il disegno campionario è stratificato, lo stimatore*

$$\widehat{\sigma}_y^2 = m_2 - \widehat{\mu}_y^2 \tag{7.39}$$

con m_2 *e* $\widehat{\mu}_y^2$ *dati rispettivamente dalle* (7.36), (7.38), *è uno stimatore corretto di* σ_y^2.

Lo stimatore (7.39) presenta una caratteristica negativa, che vale la pena mettere in rilievo. In alcuni casi, può assumere valori negativi. Si tratta di un evidente difetto, in quanto il parametro da stimare, σ_y^2, è non negativo. L'usare uno stimatore che, seppur con piccola probabilità, può assumere anche valori minori di zero è un po' un controsenso. Questo aspetto negativo è generato dal voler costruire uno stimatore corretto. In tal caso la condizione di non distorsione dello stimatore ha come conseguenza che esso (sia pure con probabilità presumibilmente "piccola") può anche assumere valori negativi.

Una possibile alternativa allo stimatore (7.39) si può costruire osservando che la varianza della popolazione, σ_y^2, si può scrivere come:

$$\sigma_y^2 = \frac{1}{N}\sum_{g=1}^{M}\sum_{i=1}^{N_g}(y_{gi} - \mu_y)^2 \tag{7.40}$$

$$= \sum_{g=1}^{M} w_g \left\{ \frac{1}{N_g}\sum_{i=1}^{N_g}(y_{gi} - \mu_y)^2 \right\}. \tag{7.41}$$

Ora, la media μ_y che compare nella (7.41) si può stimare con $\widehat{\mu}_{str}$. Inoltre, la quantità $\sum(y_{gi} - \mu_y)^2/N_g$ che compare entro parentesi nella (7.41) si può stimare con il suo "corrispondente campionario", pari a:

$$\frac{1}{n_g}\sum_{i\in\mathbf{s}_g}(y_{gi} - \widehat{\mu}_{str})^2, \quad g = 1, \ldots, M.$$

Ne segue che come stimatore intuitivo della varianza σ_y^2 della popolazione si può considerare il seguente:

$$\widehat{\sigma}_{str\,y}^2 = \sum_{g=1}^{M} w_g \left\{ \frac{1}{n_g}\sum_{i\in\mathbf{s}_g}(y_{gi} - \widehat{\mu}_{str})^2 \right\}. \tag{7.42}$$

Lo stimatore (7.42) è *distorto*. D'altra parte, esso è sempre *non negativo*, per cui in diverse circostanze si fa preferire allo stimatore (7.39).

Esercizi

7.1. Provare la relazione (7.12).

Suggerimento. Tenere conto che

$$S_y^2 = \frac{N}{N-1}\left\{\sum_{g=1}^{M} w_g \frac{N_g-1}{N_g} S_{yg}^2 + \sum_{g=1}^{M} w_g(\mu_{yg}-\mu_y)^2\right\},$$

e che

$$\frac{N}{N-1}\sum_{g=1}^{M} w_g \frac{N_g-1}{N_g} S_{yg}^2 = \left(1+\frac{1}{N-1}\right)\sum_{g=1}^{M} w_g \left(1-\frac{1}{N_g}\right) S_{yg}^2$$
$$= \sum_{g=1}^{M} w_g S_{yg}^2 + \frac{1}{N-1}\sum_{g=1}^{M} w_g \frac{N_g-1}{N_g} S_{yg}^2 - \frac{1}{N-1}\sum_{g=1}^{M} w_g \frac{N-1}{N_g} S_{yg}^2.$$

7.2. Provare che se $c_1 = \cdots = c_M$ la (7.26) si riduce alla (7.17).

7.3. Una popolazione è suddivisa in quattro strati, rispettivamente di numerosità $N_1 = 500$, $N_2 = 1000$, $N_3 = 1200$, $N_4 = 150$. Le medie di strato μ_{yg} sono incognite, ma si può ritenere, in prima approssimazione, che $\mu_{y2} = 2\,\mu_{y1}$, $\mu_{y3} = 3\,\mu_{y1}$, e $\mu_{y4} = 5\,\mu_{y1}$. Anche le varianze di strato sono incognite ma, a titolo di prima approssimazione si può assumere che siano proporzionali alle medie di strato: $S_{yg} = cost\,\mu_{yg}$, $g = 1, \ldots, 4$, essendo *cost* un'opportuna costante. Determinare l'allocazione di Neyman per una numerosità campionaria $n = 60$.

7.4. Nell'esercizio precedente si assuma che il costo di osservazione di un'unità campionaria sia di 13 Euro negli strati 1 e 2, di 5 Euro nello strato 3, e di 8 Euro nello strato 4. Assumendo un costo fisso iniziale nullo ($c_0 = 0$), determinare l'allocazione ottima quando per effettuare la rilevazione è stanziato un *budget* di 350 Euro.

7.5. Le aziende di un settore industriale sono stratificate sulla base del numero di addetti, secondo lo schema seguente. Per ogni strato sono anche riportati la media μ_{yg} e la deviazione *standard* corretta S_{yg} del fatturato annuo delle aziende (in migliaia di Euro).

Addetti	*Aziende*	μ_{yg}	S_{yg}
1–5	890	1700	38
6–20	406	50000	225
21–100	91	190000	450
101–	18	1200000	1700

Confrontare, per un campione di numerosità $n = 120$, la varianza dello stimatore $\widehat{\mu}_{str}$ con allocazione uniforme, proporzionale, e di Neyman.

7.6. Nella formula della varianza con allocazione di Neyman, il termine $\left(\sum_g w_g\, S_{yg}\right)^2$ è più piccolo di $\sum_g w_g\, S_{yg}^2$ (perché?). Pertanto, per n grande di ha che $V(\widehat{\mu}_{str};\ Ney) < 0$, un fatto privo di senso. A cosa è dovuto questo fatto?

7.7. Si consideri una popolazione finita, e si supponga di selezionare da essa:

- un campione stratificato proporzionale di $n = 100$ unità;
- un campione stratificato di Neyman di $n = 100$ unità.

Sulla base dei due campioni si costruiscono due intervalli di confidenza per μ_y, entrambi di livello $1 - \alpha$. Quale dei due ci si aspetta che abbia lunghezza minore?

7.8. Si consideri la popolazione di 1570 studenti dell'Esempio 7.1. Calcolare, per una numerosità campionaria $n = 100$, la varianza di $\widehat{\mu}_{str}$ nei due casi di allocazione proporzionale e di Neyman.

7.9. Per la popolazione di 1570 studenti dell'Esempio 7.1 determinare, nel caso di allocazione proporzionale, la numerosità campionaria necessaria affinché l'errore assoluto di stima $|\widehat{\mu}_{str} - \mu_y|$ sia maggiore di 2 cm. con probabilità pari a 0.1.

7.10. Da una popolazione di N unità suddivisa in M strati rispettivamente di numerosità $N_1, \ldots, N_M$, si seleziona un campione stratificato con numerosità camponarie di strato $n_1, \ldots, n_M$. Come stimatore di μ_y si consideri il seguente:

$$\widehat{\mu} = \sum_{g=1}^{M} c_g\, \overline{y}_g$$

essendo $c_1, \ldots, c_M$ delle costanti arbitrarie.

a. Calcolare $E[\widehat{\mu}]$.
b. Provare che $\widehat{\mu}$ è corretto se e solo se $c_1 = w_1, \ldots, c_m = w_M$.
c. Calcolare l'errore quadratico medio di $\widehat{\mu}$.

8

Disegno campionario stratificato II

8.1 Stratificazione ottimale: aspetti introduttivi

Un problema molto importante, alla base dell'utilizzo del disegno stratificato, è la costruzione degli strati. Alcuni principî fondamentali sono già stati messi in evidenza nella Sezione 7.6. Il punto chiave è che gli strati dovrebbero essere quanto più possibile omogenei, ossia formati da unità con modalità simili del carattere di interesse $\mathcal{Y}$. L'obiettivo di questa sezione è quello di fornire alcuni approfondimenti su questo importante problema. Ovviamente, un problema connesso è quello della scelta del numero degli strati. Va da sé, infatti, che quanto più numerosi sono gli strati, tanto maggiore sarà la loro omogeneità. Per semplificare le cose, nella presente sezione assumeremo *fissato* il numero M di strati. Problemi legati alla scelta di M saranno discussi nella Sezione 8.2.

Lo schema di esposizione che seguiremo è semplice. Inizieremo con assunzioni estremamente restrittive, utili non tanto per la loro immediata utilità applicativa, quanto perché dai risultati che si ottengono si hanno indicazioni utili sulle modifiche da apportare per applicare in pratica i risultati stessi.

Una volta stabilito il numero di strati in cui suddividere la popolazione, l'idea di base per costruire gli strati stessi è molto semplice: essi dovrebbero rendere minima la varianza dello stimatore $\widehat{\mu}_{str}$. In particolare, se si usa l'allocazione di Neyman la suddivisione della popolazione in M strati dovrebbe essere effettuata in modo da rendere minima la

$$V(\widehat{\mu}_{str};\ Ney) = \frac{1}{n}\left(\sum_{g=1}^{M} w_g\, S_{yg}\right)^2 - \frac{1}{N}\sum_{g=1}^{M} w_g\, S_{yg}^2.$$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_8,

Se N è sufficientemente grande, il termine $\sum_g w_g\, S_{yg}^2/N$ si può trascurare, per cui minimizzare $V(\widehat{\mu}_{str};\ Ney)$ equivale (in via approssimata) a minimizzare

$$\frac{1}{n}\left(\sum_g w_g\, S_{yg}\right)^2$$

il che, a sua volta, equivale a minimizzare $\sum_g w_g\, S_{yg}$.

*8.1.1 Teoria di base: le equazioni di Dalenius**

Per iniziare, assumiamo che il carattere di interesse $\mathcal{Y}$ sia noto, e che abbia distribuzione assolutamente continua sull'intervallo $(y_{min},\ y_{max})$, essendo y_{min} la più piccola modalità di $\mathcal{Y}$ e y_{max} la più grande. Sia $f_Y(y)$ la funzione di densità di $\mathcal{Y}$. In queste condizioni, suddividere la popolazione in strati significa (come detto nella Sezione 7.6) determinare $M-1$ "punti di taglio" $l_1, \ldots, l_{M-1}$, e nel definire gli strati nel modo seguente:

- strato 1: insieme delle unità i con modalità $y_{min} \leqslant y_i \leqslant l_1$;
- strato 2: insieme delle unità i con modalità $l_1 < y_i \leqslant l_2$;
  ...
- strato M: insieme delle unità i con modalità $l_{M-1} < y_i \leqslant y_{max}$.

I punti $l_1, \ldots, l_{M-1}$ vanno determinati in modo da rendere minima la

$$\sum_{g=1}^{M} w_g\, S_{yg}. \tag{8.1}$$

Per comodità di notazione poniamo $l_0 = y_{min}$, $l_M = y_{max}$. Valgono le seguenti, fondamentali relazioni:

$$\begin{aligned}
w_g &= \int_{l_{g-1}}^{l_g} f_Y(y)\, dy\,, \quad g = 1, \ldots, M;\\
\mu_{yg} &= \frac{1}{w_g}\int_{l_{g-1}}^{l_g} y\, f_Y(y)\, dy\,, \quad g = 1, \ldots, M;\\
S_{yg}^2 &= \frac{1}{w_g}\int_{l_{g-1}}^{l_g} (y-\mu_{yg})^2\, f_Y(y)\, dy\\
&= \frac{1}{w_g}\int_{l_{g-1}}^{l_g} y^2\, f_Y(y)\, dy - \mu_{yg}^2\,, \quad g = 1, \ldots, M.
\end{aligned}$$

Proposizione 8.1. *I valori di $l_1, \ldots, l_{M-1}$ che minimizzano la* (8.1) *soddisfano le equazioni*

$$\frac{S_{yg}^2 + (l_g - \mu_{yg})^2}{S_{yg}} = \frac{S_{y\,g+1}^2 + (l_{g+1} - \mu_{y\,g+1})^2}{S_{y\,g+1}}\,, \quad g = 1, \ldots, M-1. \tag{8.2}$$

Dimostrazione. I punti $l_1, \ldots, l_{M-1}$ si ricavano risolvendo le equazioni:

$$\frac{\partial}{\partial l_g}\left(\sum_{h=1}^{M} w_h\, S_{yh}\right) = 0, \quad g = 1, \ldots, M-1. \tag{8.3}$$

Ora, nella somma che compare in (8.3) solo $w_g\, S_{yg}$ e $w_{g+1}\, S_{y\,g+1}$ dipendono da l_g, per cui si ha

$$\frac{\partial}{\partial l_g}\left(\sum_{h=1}^{M} w_h\, S_{yh}\right) = \frac{\partial w_g\, S_{yg}}{\partial l_g} + \frac{\partial w_{g+1}\, S_{y\,g+1}}{\partial l_g}$$

e le (8.3) diventano

$$\frac{\partial w_g\, S_{yg}}{\partial l_g} + \frac{\partial w_{g+1}\, S_{y\,g+1}}{\partial l_g} = 0\,, \quad g = 1, \ldots, M-1. \tag{8.4}$$

Per calcolare le derivate parziali in (8.4) iniziamo con il calcolare le derivate (rispetto a l_g) di w_g, $w_g\,\mu_{yg}$, $w_g\, S_{yg}^2$, pari a:

$$\frac{\partial\, w_g}{\partial\, l_g} = \frac{\partial}{\partial\, l_g}\left\{\int_{l_{g-1}}^{l_g} f_Y(y)\, dy\right\} = f_y(l_g); \tag{8.5}$$

$$\begin{aligned}
\frac{\partial\, w_g\, \mu_{yg}}{\partial\, l_g} &= \frac{\partial}{\partial\, l_g}\left\{\int_{l_{g-1}}^{l_g} y f_Y(y)\, dy\right\} = l_g\, f_y(l_g);\\
\frac{\partial\, w_g\, S_{yg}^2}{\partial\, l_g} &= \frac{\partial}{\partial\, l_g}\left\{\int_{l_{g-1}}^{l_g} y^2 f_Y(y)\, dy - \frac{1}{w_g}\,(w_g\,\mu_{yg})^2\right\}\\
&= \frac{\partial}{\partial\, l_g}\left\{\int_{l_{g-1}}^{l_g} y^2 f_Y(y)\, dy\right\} - \frac{1}{w_g^2}\left\{w_g\,\frac{\partial\,(w_g\,\mu_{yg})^2}{\partial\, l_g} - (w_g\,\mu_{yg})^2\,\frac{\partial\, w_g}{\partial\, l_g}\right\}\\
&= l_g^2\, f_Y(l_g) - \frac{1}{w_g^2}\left\{2\, w_g^2\,\mu_{yg}\,\frac{\partial\, w_g\,\mu_{yg}}{\partial\, l_g} - w_g^2\,\mu_{yg}^2\, f_Y(l_g)\right\}\\
&= l_g^2\, f_Y(l_g) - 2\, l_g\,\mu_{yg}\, f_Y(l_g) + \mu_{yg}^2\, f_Y(l_g)\\
&= f_Y(l_g)\,(l_g - \mu_{yg})^2.
\end{aligned} \tag{8.6}$$

Tenendo poi conto che

$$\frac{\partial\,(w_g\, S_{yg})^2}{\partial\, l_g} = 2\, w_g\, S_{yg}\,\frac{\partial\, w_g\, S_{yg}}{\partial\, l_g} \quad \textit{da cui segue che}$$

$$\frac{\partial\, w_g\, S_{yg}}{\partial\, l_g} = \frac{1}{2\, w_g\, S_{yg}}\,\frac{\partial\,(w_g\, S_{yg})^2}{\partial\, l_g}$$

e che

$$\frac{\partial\,(w_g\, S_{yg})^2}{\partial\, l_g} = w_g\,\frac{\partial\, w_g\, S_{yg}^2}{\partial\, l_g} + w_g\, S_{yg}^2\,\frac{\partial\, w_g}{\partial\, l_g}$$

dalla (8.6) si ottiene

$$\frac{\partial\, w_g\, S_{yg}}{\partial\, l_g} = \frac{1}{2\, w_g\, S_{yg}} \left\{ w_g\, f_Y(l_g)\, (S^2_{yg} + (l_g - \mu_{yg})^2) + w_g\, S^2_{yg}\, f_Y(l_g) \right\}$$
$$= \frac{1}{2\, S_{yg}}\, f_Y(l_g)\, \{ S^2_{yg} + (l_g - \mu_{yg})^2 \}. \tag{8.7}$$

Nello stesso modo si prova che

$$\frac{\partial\, w_{g+1}\, S_{y\,g+1}}{\partial\, l_g} = -\frac{1}{2\, S_{y\,g+1}}\, f_Y(l_g)\, \left\{ S^2_{y\,g+1} + (l_g - \mu_{y\,g+1})^2 \right\}. \tag{8.8}$$

Dalle (8.4), (8.7), (8.8) si ottengono subito le (8.2). □

Le equazioni (8.2) sono essenzialmente dovute a Dalenius (1950).

*8.1.2 Equazioni di Dalenius basate su un carattere ausiliario**

Le equazioni di Dalenius (8.2) non hanno, di fatto, nessuna applicabilità diretta, in quanto presuppongono che la distribuzione (di frequenza) del carattere di interesse $\mathcal{Y}$ (nella popolazione di riferimento) sia nota e assolutamente continua (ossia dotata di densità). Dei due punti deboli il primo è di gran lunga il più grave. Se la distribuzione di $\mathcal{Y}$ è nota, perché ricorrere ad una rilevazione campionaria? D'altra parte, la Proposizione 8.1 si rivela molto utile sul piano applicativo quando è noto un carattere ausiliario $\mathcal{X}$, fortemente correlato con $\mathcal{Y}$. In questo caso si può pensare di stratificare le unità della popolazione sulla base delle modalità $x_1, \ldots, x_N$ di $\mathcal{X}$, usato in qualche modo come "surrogato" di $\mathcal{Y}$. Questo significa, in sostanza, cercare di definire gli strati in modo da rendere minima la quantità

$$\sum_{g=1}^{M} w_g\, S_{xg}. \tag{8.9}$$

L'idea di base dietro questo procedimento, dettata direttamente dall'intuizione, è che la correlazione (che si spera forte) tra $\mathcal{X}$ e $\mathcal{Y}$ dovrebbe fare in modo che minimizzare la (8.9) porti ad una stratificazione della popolazione assai simile a quella che si avrebbe minimizzando la (8.1).

I risultati che si ottengono utilizzando un carattere ausiliario $\mathcal{X}$ sono del tutto simili a quelli già visti nella sezione precedente. Supponiamo, per semplicità di trattazione, che $\mathcal{X}$ abbia, nella popolazione di interesse, distribuzione assolutamente continua con funzione di densità $f_X(x)$ definita nell'intervallo $(x_{min},\ x_{max})$, dove x_{min} è la più piccola modalità di $\mathcal{X}$, e x_{max} la più grande. Posto $t_0 = x_{min}$, $t_M = x_{max}$, l'idea di base, ricalcata sulla precedente sezione, è quella di prendere $M-1$ punti $t_1, \cdots, t_{M-1}$ dell'intervallo $(x_{min},\ x_{max})$, e di definire gli strati come segue:

– strato 1: insieme delle unità i con modalità $t_0 \leqslant x_i \leqslant t_1$;
  ...
– strato M: insieme delle unità i con modalità $t_{(M-1)} < x_i \leqslant t_M$.

I punti $t_1, \ldots, t_{M-1}$ vanno determinati in modo da soddisfare le equazioni (8.2), ma con medie e varianze di strato riferite al carattere ausiliario $\mathcal{X}$ anziché a $\mathcal{Y}$. In altri termini, e usando un'ovvia notazione, $t_1, \ldots, t_{M-1}$ sono determinati sulla base delle relazioni

$$\frac{S_{xg}^2 + (t_g - \mu_{xg})^2}{S_{xg}} = \frac{S_{x\,g+1}^2 + (t_{g+1} - \mu_{x\,g+1})^2}{S_{x\,g+1}}, \quad g = 1, \ldots, M-1. \tag{8.10}$$

In generale, equazioni (non lineari) del tipo (8.10), (8.2) sono assai difficili da risolvere.

Nel seguito esamineremo diversi adattamenti e approssimazioni utili per le applicazioni.

*8.1.3 Regole approssimate per la stratificazione ottima**

Quanto detto nella sezione precedente mostra come sia di primaria importanza cercare soluzioni approssimate delle (8.10). Una delle più importanti è quella dovuta a Dalenius e Hodges (1959), esposta in questa sezione.

Se gli strati sono abbastanza numerosi, la densità $f_X(x)$ si può considerare all'incirca costante in ogni strato. Formalmente, si può scrivere:

$$f_X(x) \approx f_{xg} \;\; \textit{per ogni } t_{g-1} < x \leqslant t_g, \;\; g = 1, \ldots, M\,,$$

da cui:

$$\begin{aligned}
w_g &= \int_{t_{g-1}}^{t_g} f_X(x)\,dx \approx f_{xg}\,(t_g - t_{g-1}), \;\; g = 1, \ldots, M; \qquad (8.11)\\
\mu_{xg} &= \frac{1}{w_g}\int_{t_{g-1}}^{t_g} x\, f_X(x)\,dx \\
&= \frac{1}{t_g - t_{g-1}}\int_{t_{g-1}}^{t_g} x\,dx \\
&= \frac{1}{2}\,(t_g + t_{g-1}), \;\; g = 1, \ldots, M; \\
S_{xg}^2 &= \frac{1}{w_g}\int_{t_{g-1}}^{t_g} (x - \mu_{xg})^2\, f_X(x)\,dx \\
&\approx \frac{1}{t_g - t_{g-1}}\int_{t_{g-1}}^{t_g} \left(x - \frac{t_g + t_{g-1}}{2}\right)^2 dx \\
&= \frac{1}{12}\,(t_g - t_{g-1})^2, \;\; g = 1, \ldots, M; \\
S_{xg} &\approx \frac{1}{\sqrt{12}}\,(t_g - t_{g-1}), \;\; g = 1, \ldots, M. \qquad (8.12)
\end{aligned}$$

Poniamo adesso

$$D_g = \int_{t_{g-1}}^{t_g} \sqrt{f_X(x)}\, dx, \quad g = 1, \ldots, M. \tag{8.13}$$

Data la $f_X(x)$, la somma dei termini D_g è fissata, e pari a

$$K = \int_{t_0}^{t_M} \sqrt{f_X(x)}\, dx \tag{8.14}$$

in quanto

$$\sum_{g=1}^{M} D_g = \sum_{g=1}^{M} \int_{t_{g-1}}^{t_g} \sqrt{f_X(x)}\, dx = \int_{t_0}^{t_M} \sqrt{f_X(x)}\, dx = K. \tag{8.15}$$

Se $f_X(x)$, come assunto in precedenza, è all'incirca costante su tutto l'intervallo $(t_{g-1},\ t_g)$, $f_X(x) \approx f_{xg}$ per $t_{g-1} < x \leqslant t_g$, si ha

$$\begin{aligned} D_g &\approx \int_{t_{g-1}}^{t_g} \sqrt{f_{xg}}\, dx \\ &= \sqrt{f_{xg}}\,(t_g - t_{g-1})\,, \quad g = 1, \ldots, M \end{aligned}$$

e quindi, usando anche le (8.11), (8.12), valgono le relazioni

$$\begin{aligned} w_g\, S_{xg} &\approx \frac{1}{\sqrt{12}}\, f_{xg}\,(t_g - t_{g-1})^2 \\ &\approx \frac{1}{\sqrt{12}}\, D_g^2\,, \quad g = 1, \ldots, M. \end{aligned} \tag{8.16}$$

Da queste approssimazioni discende subito che la (8.9) si approssima nel modo seguente

$$\sum_{g=1}^{M} w_g\, S_{xg} \approx \frac{1}{\sqrt{12}} \sum_{g=1}^{M} D_g^2$$

e dunque minimizzare la (8.9) equivale a minimizzare la $\sum_{g=1}^{M} D_g^2$. In vista delle (8.15), (8.14), la somma delle D_g deve soddisfare il vincolo $D_1 + \cdots + D_M = K$ (K fissato). Formalmente, per determinare $t_1, \ldots, t_{M-1}$ bisogna risolvere il seguente problema di minimo vincolato:

$$\begin{cases} \textit{minimizzare}: \ \sum_{g=1}^{M} D_g^2 \\ \textit{con il vincolo}: \ \sum_{g=1}^{M} D_g = K \end{cases}. \tag{8.17}$$

Usando la solita tecnica dei moltiplicatori di Lagrange, è immediato verificare (Esercizio 8.1) che la soluzione del problema di minimo (8.17) prevede valori D_g tutti uguali: $D_1 = \cdots = D_M = K/M$. Ricordando la (8.13), ciò significa scegliere $t_1, \ldots, t_{M-1}$ in maniera tale che le quantità

$$\int_{t_{g-1}}^{t_g} \sqrt{f_X(x)}\, dx\,, \quad g = 1, \ldots, M$$

siano tutti uguali. È questa la celebre *regola cum* $\sqrt{f}$ di Dalenius e Hodges (1959). Essa si basa sull'assunzione che la distribuzione di frequenza del carattere $\mathcal{X}$ nella popolazione possieda una densità $f_X(x)$. In pratica, questo non accade mai. La popolazione di riferimento è finita, composta da N unità, e il carattere $\mathcal{X}$ non può che essere discreto. In questo caso il ruolo di densità di $\mathcal{X}$ è svolto dal suo istogramma. In pratica, l'intervallo $(x_{min},\ x_{max})$ viene suddiviso in H intervalli uguali; in generale H va scelto in modo da essere molto più grande di M (numero degli strati). Per l'intervallo h-mo, sia f_h la frequenza relativa delle modalità di $\mathcal{X}$ nell'intervallo h-mo ($h = 1, \ldots, H$). L'altezza dell'istogramma (e quindi la densità) nella classe h-ma è ovviamente $f_h/((x_{max} - x_{min})/H)$. Gli M strati vengono formati aggregando classi contigue, a partire dalla coda sinistra dell'istogramma. La regola cum $\sqrt{f}$ stabilisce che le classi consecutive vanno aggregate in modo che la somma delle $\sqrt{f_h}$ in ogni strato sia approssimativamente costante, e quindi approssimativamente uguale a $\sum_{h=1}^{H} \sqrt{f_h}/M$.

Esempio 8.1. Nel *file* `impr80.txt` sono riportati dati relativi ad una popolazione di 385 aziende manifatturiere (numero di dipendenti, fatturato, flusso di cassa, etc.) per l'anno 1980. Si tratta essenzialmente di dati reali, disponibili al sito *web ftp://elsa.berkeley.edu/users/bhhall/pub/data/*. In particolare, nei dati di questo esempio si considerano solo imprese con non più di 1500 addetti.

Supponiamo di conoscere il numero di addetti di ogni azienda. Grandezze quali fatturato, flusso di cassa, etc., presentano una correlazione abbastanza forte con il numero di dipendenti, che può essere perciò usato come carattere di stratificazione. La sua distribuzione è mostrata in Fig. 8.1, che riporta il relativo istogramma.

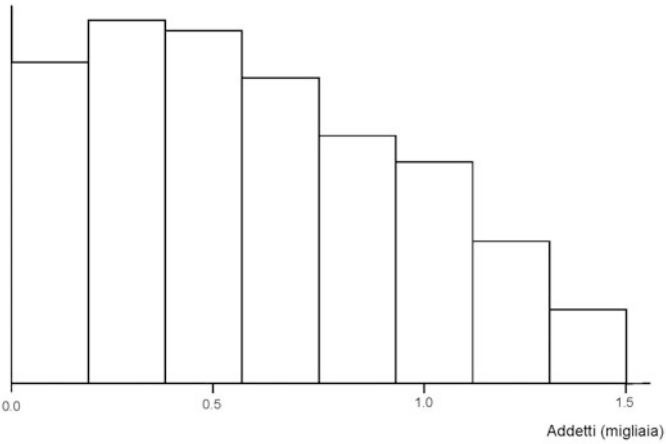


Fig. 8.1 Istogramma del numero di occupati di 385 aziende

Tabella 8.1 Stratificazione con regola cum $\sqrt{f}$

Classe occupati (migliaia)	h	*Frequenza relativa* f_h	$\sqrt{f_h}$	*Classe occupati (migliaia)*	h	*Frequenza relativa* f_h	$\sqrt{f_h}$
$0.0000 - 0.0375$	1	0.0052	0.072	$0.7500 - 0.7875$	21	0.0312	0.1765
$0.0375 - 0.0750$	2	0.0156	0.1248	$0.7875 - 0.8250$	22	0.0208	0.1441
$0.0750 - 0.1125$	3	0.0338	0.1838	$0.8250 - 0.8265$	23	0.0182	0.1348
$0.1125 - 0.1500$	4	0.0519	0.2279	$0.8265 - 0.9000$	24	0.0390	0.1974
$0.1500 - 0.1875$	5	0.0338	0.1838	$0.9000 - 0.9375$	25	0.0156	0.1248
$0.1875 - 0.2250$	6	0.0338	0.1838	$0.9375 - 0.9750$	26	0.0156	0.1248
$0.2250 - 0.2625$	7	0.0468	0.2162	$0.9750 - 1.0125$	27	0.0208	0.1441
$0.2625 - 0.3000$	8	0.0182	0.1348	$1.0125 - 1.0500$	28	0.0260	0.1612
$0.3000 - 0.3375$	9	0.0338	0.1838	$1.0500 - 1.0875$	29	0.0238	0.1529
$0.3375 - 0.3750$	10	0.0338	0.1838	$1.0875 - 1.1250$	30	0.0260	0.1612
$0.3750 - 0.4125$	11	0.0390	0.1974	$1.1250 - 1.1625$	31	0.0130	0.1140
$0.4125 - 0.4500$	12	0.0364	0.1907	$1.1625 - 1.2000$	32	0.0208	0.1441
$0.4500 - 0.4875$	13	0.0468	0.2162	$1.2000 - 1.2375$	33	0.0026	0.0510
$0.4875 - 0.5250$	14	0.0364	0.1907	$1.2375 - 1.2750$	34	0.0104	0.1019
$0.5250 - 0.5625$	15	0.0260	0.1612	$1.2750 - 1.3125$	35	0.0130	0.1140
$0.5625 - 0.6000$	16	0.0208	0.1441	$1.3125 - 1.3500$	36	0.0182	0.1348
$0.6000 - 0.6375$	17	0.0442	0.2101	$1.3500 - 1.3875$	37	0.0182	0.1348
$0.6375 - 0.6750$	18	0.0208	0.1441	$1.3875 - 1.4250$	38	0.0156	0.1248
$0.6750 - 0.7125$	19	0.0234	0.1529	$1.4250 - 1.4265$	39	0.0156	0.1248
$0.7125 - 0.7500$	20	0.0234	0.1529	$1.4265 - 1.5000$	40	0.0130	0.1140

Supponiamo di voler costruire, in totale, $M = 4$ strati. In Tabella 8.1 sono riportate le grandezze che intervengono nella costruzione effettiva degli strati, avendo diviso l'intervallo $[0, 1.5]$ in $H = 40$ classi di occupati ciascuna di ampiezza 0.0375. Per convenzione, ogni classe include il suo estremo destro ma non quello sinistro. Con una certa approssimazione (non molta, per la verità), il numero di occupati in migliaia viene trattato come una variabile continua pur non essendolo realmente.

Essendo $\sum_{h=1}^{40} \sqrt{f_h} = 6.1351$, la regola cum $\sqrt{f}$ prevede che la somma delle $\sqrt{f_h}$ in ogni strato sia (approssimativamente) uguale a $6.1351/4 = 1.5338$.

Dalla (8.1) si deduce facilmente che

$$\sum_{h=1}^{9} \sqrt{f_h} = 1.5114, \quad \sum_{h=10}^{17} \sqrt{f_h} = 1.4942, \quad \sum_{h=18}^{27} \sqrt{f_h} = 1.4964, \sum_{h=28}^{40} \sqrt{f_h} = 1.6335$$

e quindi i 4 strati che si ottengono dalla regola cum $\sqrt{f}$ sono i seguenti:

- strato 1: insieme delle imprese fino a 337 dipendenti;
- strato 2: insieme delle imprese con più di 337 e non oltre 637 dipendenti;
- strato 3: insieme delle imprese con più di 637 e non oltre 1012 dipendenti;
- strato 4: insieme delle imprese con più di 1012 e non oltre 1500 dipendenti.

□

La regola cum $\sqrt{f}$ stabilisce in sostanza che $t_1, \ldots, t_{M-1}$ vanno scelti in modo che le grandezze

$$\int_{t_{g-1}}^{t_g} \sqrt{f_X(x)}\, dx \approx \sqrt{f_{xg}}\,(t_g - t_{g-1})$$

siano costanti, e quindi che siano costanti le $f_{xg}\,(t_g - t_{g-1})^2$. Usando la (8.16), questo equivale a richiedere che i prodotti

$$w_g\, S_{xg}, \quad g = 1, \ldots, M$$

siano costanti.

In letteratura vi sono diverse metodi approssimati di stratificazione ottimale, alternative alla regola cum $\sqrt{f}$. Una delle più semplici è la regola di Ekman (1959), la quale stabilisce di scegliere $t_1, \ldots, t_{M-1}$ in modo che siano costanti le grandezze

$$w_g\,(t_g - t_{g-1}), \quad g = 1, \ldots, M. \tag{8.18}$$

Poiché la distribuzione del carattere $\mathcal{X}$ è discreta, in genere le (8.18) potranno essere solo approssimativamente costanti. Inoltre, la regola di Ekman equivale a rendere (approssimativamente) costanti le quantità $N_g(t_g - t_{g-1})$.

Se $f_X(x) \approx f_{xg}$ si ha, in vista delle (8.11), (8.12), (8.16), $w_g\,(t_g - t_{g-1}) \approx w_g\, S_{xg}/\sqrt{12}$, per cui anche la regola di Ekman, in via largamente approssimata, può essere considerata pressoché equivalente alla regola cum $\sqrt{f}$.

Per altri contributi alla teoria della stratificazione ottima si rinvia ai lavori di Serfling (1968), Singh (1971), Hedlin (2000). Confronti tra i diversi metodi di stratificazione basati su dati reali sono in Cochran (1961) Hess *e altri* (1966). Da questi lavoro emerge come la regola di Ekman abbia un comportamento lievemente migliore della cum $\sqrt{f}$. Le buone caratteristiche della regola di Ekman sono anche confermate in Murthy (1967), anche se è da rimarcare che nel confronto effettuato da questo autore non viene considerata la regola cum $\sqrt{f}$. Si tratta comunque di conclusioni di portata non decisiva, in quanto basati solo su pochi insiemi di dati reali.

Contrariamente a quel che accade per la regola cum $\sqrt{f}$, la regola di Ekman è molto più difficile da applicare in pratica. Per la sua implementazione è quasi sempre necessario procedere per via numerica. Qui di seguito è brevemente indicato un semplice algoritmo per il calcolo di $t_1, \ldots, t_{M-1}$. Esso richiede la specificazione *a priori* di un numero $\delta > 0$ "piccolo" (tolleranza), nonché di un valore iniziale $x_{min} \leqslant \tau \leqslant x_{min}$ per t_1. Nel seguito, indicheremo anche con $N(t)$ il numero di unità della popolazione con modalità $x_i \leqslant t$, e porremo $F(t) = N(t)/N$.

- *Passo* 0. Inizializzazione. Porre $\tau_1 = \tau$, $l = x_{min}$, $u = x_{max}$. Andare al Passo 1.
- *Passo* 1. Calcolare $E_1 = (\tau_1 - x_{min}) \times F(\tau_1)$. Andare al Passo 2.

- *Passo* 2. Determinare $\tau_2, \ldots, \tau_{M-1}$ tali che, posto $E_g = (\tau_g - \tau_{g-1}) \times (F(\tau_g) - F(\tau_{g-1}))$, sia $|E_g - E_1| < \delta$ per ogni $g = 2, \ldots, M-1$. Se questi valori non esistono (perché si trova $\tau_g \geqslant x_{max}$ per qualche g), modificare τ_1 come $\tau_1 = \tau_1/2$, porre $u = \tau_1$, e tornare al Passo 1; altrimenti, andare al Passo 3.
- *Passo* 3. Calcolare $E_M = (x_{max} - \tau_{M-1}) \times (1 - F(\tau_{M-1}))$. Se $E_M - E_1 > \delta$ andare al Passo 4. Se $E_1 - E_M > \delta$ andare al Passo 5. Se $|E_M - E_1| \leqslant \delta$ andare al Passo 6.
- *Passo* 4. Porre $l = \tau_1$, e modificare τ_1 ponendo $\tau_1 = (l + u)/2$. Andare al Passo 1.
- *Passo* 5. Porre $u = \tau_1$ e modificare τ_1 ponendo $\tau_1 = (l + u)/2$. Andare al Passo 1.
- *Passo* 6. Stop. Porre $t_1 = \tau_1, \ldots, t_{M-1} = \tau_{M-1}$.

Esempio 8.2. Consideriamo la popolazione di 385 imprese del *file* `impr80.txt` (Esempio 8.1). L'obiettivo è di costruire $M = 4$ strati usando la regola di Ekman. Per semplicità di calcolo, e similmente all'Esempio 8.1, si considerano $H = 40$ classi di occupati ciascuna di ampiezza 0.0375. Il termine δ è qui scelto pari a 0.015, mentre il valore iniziale τ è posto pari a 0.675. Usando l'algoritmo precedentemente illustrato, si ha che $t_1 = 0.338$, $t_2 = 0.675$, $t_3 = 1.05$, a cui corrispondono valori $w_g\,(t_g - t_{g-1})$ del tipo:

$$w_1\,(t_1 - t_0) = 0.09,\ w_2\,(t_2 - t_1) = 0.102,\ w_3\,(t_3 - t_2) = 0.087,$$
$$w_4\,(t_4 - t_3) = 0.085.$$

I quattro strati che si ottengono dalla regola cum $\sqrt{f}$ sono qui di seguito riportati. Si tratta di strati quasi uguali a quelli ottenuti nell'Esempio 8.1 usando la regola cum $\sqrt{f}$:

- strato 1: insieme delle imprese fino a 338 dipendenti;
- strato 2: insieme delle imprese con più di 338 e non oltre 675 dipendenti;
- strato 3: insieme delle imprese più di 675 e non oltre 1050 dipendenti;
- strato 4: insieme delle imprese più di 1050 e non oltre 1500 dipendenti.

□

Dopo che la popolazione è stata suddivisa in strati, rimane da risolvere il problema dell'allocazione delle n unità campionarie negli M strati. Un approccio molto semplice consiste nell'usare l'allocazione di Neyman in cui le incognite S_{yg}^2 sono sostituite con le corrispondenti varianze di strato S_{xg}^2 del carattere $\mathcal{X}$. Formalmente, dallo strato g si seleziona (mediante disegno ssr) un numero di unità pari a:

$$n_g = n\,\frac{w_g\,S_{xg}}{\sum_{h=1}^{M} w_h\,S_{xh}},\quad g = 1, \ldots, M.$$

Questo tipo di approccio fornisce risultati tanto migliori quanto più le S_{xg} sono dei "buoni sostituti" delle S_{yg}. Questo, in genere, accade se i due caratteri $\mathcal{X}$,

$\mathcal{Y}$ sono fortemente correlati. In alternativa all'allocazione di Neyman basata sulle S_{xg} si può usare la procedura in due fasi descritta nella Sezione 7.4.3.

In alcune circostanze si hanno popolazioni altamente asimmetriche, in cui molte unità presentano modalità abbastanza piccole, mentre poche unità hanno modalità molto grandi, e quindi sono particolarmente importanti nel determinare la media (o l'ammontare) del carattere oggetto di interesse. Un esempio molto comune è quello del campionamento di imprese delle quali sia noto *a priori* il numero di occupati, che quindi può essere usato come carattere di stratificazione (cfr. Esempio 8.1). Accade spesso che la popolazione sia composta di una miriade di imprese medio-piccole, e di poche imprese grandi o grandissime, determinanti per lo studio dell'intera popolazione. In questi casi la procedura *standard* consiste nel suddividere la popolazione in M strati, tali che:

- lo strato M-mo è formato dalle unità "più grandi", che vengono tutte incluse nel campione (unità "auto-rappresentative");
- i restanti $M-1$ strati sono formati dalle restanti unità, e da ognuno di essi viene selezionato un campione ssr.

Si tratta, in pratica, di un disegno stratificato in cui le unità di uno strato vengono osservate tutte. Il problema della costruzione di strati ottimale, in questo ambito, è affrontato in Lavallée e Hidiroglou (1988), Hidiroglou e Srinath (1993), Rivest (2002). Per un'applicazione (una delle molte, in verità) si rinvia al lavoro di Slanta e Krenzke (1990).

In chiusura di sezione è da sottolineare che un elemento decisivo nella stratificazione basata su un carattere ausiliario è il tipo di relazione di (inter)dipendenza che esso possiede con il carattere di interesse. Questo punto è stato più volte sfiorato nelle sezioni precedenti, anche se non si è mai fatto esplicito riferimento ad un *modello di superpopolazione* che formalizzi tale (inter)dipendenza . In questo tipo di modelli (a cui si è già fatto brevemente riferimento nel Capitolo 5) si assume che le y_i non siano semplici numeri, ma realizzazioni di variabili aleatorie Y_i, legate alle x_i da un modello statistico di (inter)dipendenza. In realtà, una parte molto importante della moderna teoria della stratificazione si basa proprio su modelli di questo tipo (Särndal *e altri* (1993)). Dato il livello molto elementare di questa parte non proseguiamo oltre in questa direzione.

8.2 Considerazioni sul numero degli strati

8.2.1 Aspetti di base

Una questione di notevole importanza, ed a cui si è dedicato assai poco spazio nel capitolo precedente, è quella relativa al numero di strati da costruire . In effetti, nella costruzione degli strati ottimali svolta nelle sezioni precedenti si è sempre assunto dato il numero M di strati.

Intuitivamente, un aumento del numero M di strati dovrebbe permettere la costruzione di strati più omogenei dal punto di vista delle modalità y, e quindi dovrebbe portare ad una maggiore efficienza complessiva dello stimatore $\widehat{\mu}_{str}$. Da questo punto di vista, pertanto, converrebbe prendere un numero di strati quanto più possibile elevato, compatibilmente con la numerosità campionaria totale n. L'obiettivo della presente sezione è quello di fornire qualche precisazione, basata più sull'intuizione che sul formalismo, di questa affermazione.

Gli strati, come detto nelle sezioni precedenti, sono costruiti prendendo unità con modalità "vicine" del carattere di stratificazione $\mathcal{X}$. La varianza dello stimatore $\widehat{\mu}_{str}$ dipende essenzialmente dalla varianza di $\mathcal{Y}$ negli strati, cioè dalla "disomogeneità" delle y_i all'interno degli strati. Questa, a sua volta, dipende da due fattori: la variabilità delle modalità di $\mathcal{X}$ all'interno degli strati, e la variabilità delle y_i corrispondenti ad ogni specifica modalità di $\mathcal{X}$. In altre parole, due unità di uno stesso strato possono avere differenti valori y_i sia perché ad esse corrispondono differenti modalità di $\mathcal{X}$ (variabilità di $\mathcal{Y}$ dovuta a quella di $\mathcal{X}$), sia perché esse possono essere diverse anche a parità di modalità di $\mathcal{X}$ (variabilità *residua* di $\mathcal{Y}$).

Se si accresce il numero degli strati, tenderà a diminuire il numero di modalità distinte di $\mathcal{X}$ in ciascuno strato. Quindi, tenderà a ridursi la varianza di $\mathcal{X}$ negli strati. Ciò che resta sostanzialmente invariata, invece, è la varianza residua di $\mathcal{Y}$. Pertanto, l'aumentare il numero di strati diminuisce la variabilità di $\mathcal{Y}$ all'interno degli strati quasi soltanto perché riduce la parte di variabilità di $\mathcal{Y}$ dovuta a $\mathcal{X}$; la varianza residua di $\mathcal{Y}$, invece, a meno di casi specialissimi non viene ridotta accrescendo il numero degli strati. Questo spiega (o almeno dovrebbe spiegare) perché in parecchi casi non conviene prendere un numero di strati molto alto. Quando si accresce il numero di strati oltre un certo limite non si ottengono significativi vantaggi, perché la parte di varianza di strato di $\mathcal{Y}$ dovuta a $\mathcal{X}$, già molto piccola, diminuisce di pochissimo. Diviene invece preponderante la variabilità residua, che però non è intaccata in misura significativa da incrementi del numero di strati.

*8.2.2 Qualche risultato teorico**

Le affermazioni precedenti posso essere precisate un po' meglio, anche se è necessario complicare non poco la notazione. Per semplificare le cose supponiamo, come in precedenza, che il carattere di interesse $\mathcal{Y}$ e quello di stratificazione $\mathcal{X}$ possiedano densità, rispettivamente $f_Y(y)$ e $f_X(x)$. Questo ovviamente non è vero nella pratica, ma serve a semplificare la notazione. Essendo la popolazione di riferimento finita, i caratteri $\mathcal{Y}$, $\mathcal{X}$ sono discreti. Per riportarsi a questo caso basta sostituire densità con frequenze relative e integrali con somme.

Indichiamo con $f_{Y|X}(y\,|x)$ la densità di $\mathcal{Y}$ condizionata al valore x di $\mathcal{X}$. Detto in termini assai rozzi, questa è la densità di $\mathcal{Y}$ quando ci si restringe alle sole unità la cui modalità di $\mathcal{X}$ è x. Sia inoltre $\mu_y(x)$ la media di $\mathcal{Y}$

condizionatamente alla modalità x di $\mathcal{X}$:

$$\mu_y(x) = \int y\, f_{Y|X}(y\,|x)\, dy$$

e sia $S_y^2(x)$ la varianza di $\mathcal{Y}$ subordinatamente a x:

$$S_y^2(x) = \int (y - \mu_y(x))^2 f_{Y|X}(y\,|x)\, dy.$$

Come ben noto, valgono le due relazioni

$$\mu_y = \int \mu_y(x)\, f_X(x)\, dx, \tag{8.19}$$

$$S_y^2 = \int (\mu_y(x) - \mu_y)^2\, f_X(x)\, dx + \int S_y^2(x)\, f_X(x)\, dx. \tag{8.20}$$

Gli strati sono formati come descritto nelle Sezioni 8.1.2, 8.1.3, di cui si lascia invariata la notazione. Lo strato g-mo è formato dalle unità con modalità x nell'intervallo di estremi t_{g-1}, t_g $(g = 1, \ldots, M)$. Se restringiamo $\mathcal{X}$ allo strato g-mo, la sua densità diventa pari (sempre usando la notazione della Sezione 8.1.3) $f_X(x)/w_g$. Dalle (8.19), (8.20), "ristrette" allo strato g-mo si ricavano pertanto le seguenti ulteriori relazioni:

$$\mu_{yg} = \frac{1}{w_g} \int_{t_{g-1}}^{t_g} \mu_y(x)\, f_X(x)\, dx, \tag{8.21}$$

$$S_{yg}^2 = \frac{1}{w_g} \int_{t_{g-1}}^{t_g} (\mu_y(x) - \mu_{yg})^2\, f_X(x)\, dx + \frac{1}{w_g} \int_{t_{g-1}}^{t_g} S_y^2(x)\, f_X(x)\, dx. \tag{8.22}$$

Per ottenere risultati espliciti sono necessarie a questo punto alcune assunzioni supplementari, qui di seguito riportate:

- gli M strati hanno tutti la stessa ampiezza: $t_g - t_{g-1} = (x_{max} - x_{min})/M$, $g = 1, \ldots, M$;
- il disegno campionario è di tipo stratificato proporzionale: $n_g = n\, w_g$, $g = 1, \ldots, M$;
- la funzione di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ è lineare: $\mu_y(x) = \alpha_{y/x} + \beta_{y/x}\, x$;
- la varianza di $\mathcal{Y}$ condizionatamente a x è costante (omoscedasticità): $S_y^2(x) = S_e^2$.

Con le ipotesi fatte, le (8.19), (8.20), (8.21), (8.22) si riscrivono come:

$$\begin{aligned} \mu_y &= \alpha_{y/x} + \beta_{y/x}\, \mu_x \\ \mu_{yg} &= \frac{1}{w_g} \int_{t_{g-1}}^{t_g} (\alpha_{y/x} + \beta_{y/x}\, x)\, f_X(x)\, dx \\ &= \alpha_{y/x} + \beta_{y/x}\, \mu_{xg} \end{aligned}$$

$$
\begin{aligned}
S^2_{yg} &= \frac{1}{w_g}\int_{t_{g-1}}^{t_g} \{\alpha_{y/x} + \beta_{y/x}\, x - (\alpha_{y/x} + \beta_{y/x}\, \mu_x)\}^2 \, f_X(x)\, dx \\
&\quad + \frac{1}{w_g}\int_{t_{g-1}}^{t_g} S^2_e \, f_X(x)\, dx \\
&= \beta^2_{y/x} \frac{1}{w_g}\int_{t_{g-1}}^{t_g} (x - \mu_{xg})^2 \, f_X(x)\, dx + S^2_e \, \frac{1}{w_g}\int_{t_{g-1}}^{t_g} f_X(x)\, dx \\
&= \beta^2_{y/x}\, S^2_{xg} + S^2_e
\end{aligned}
$$

essendo μ_{xg} e S^2_{xg} rispettivamente la media e la varianza di $\mathcal{X}$ nello strato g-mo. Ne consegue che la varianza dello stimatore $\widehat{\mu}_{str}$ risulta pari a

$$
\begin{aligned}
V(\widehat{\mu}_{str}) &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{g=1}^{M} w_g S^2_{yg} \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{g=1}^{M} w_g \left(\beta^2_{y/x}\, S^2_{xg} + S^2_e\right) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\beta^2_{y/x} \sum_{g=1}^{M} w_g\, S^2_{xg} + S^2_e\right\}. \qquad (8.23)
\end{aligned}
$$

La (8.23) mette in evidenza che solo il termine $\sum_g w_g\, S^2_{xg}$ è effettivamente "sensibile" alla stratificazione, mentre il termine S^2_e non è in alcun modo intaccato dall'aver suddiviso la popolazione in strati. Non è neanche difficile vedere in che modo $\sum_g w_g\, S^2_{xg}$ diminuisce al crescere del numero di strati. Come ben noto, la varianza di un carattere è non superiore a un quarto del suo campo di variazione al quadrato. Questo significa che in ogni strato vale la disuguaglianza $S^2_{xg} \leqslant (t_g - t_{g-1})^2/4$. Sfruttando anche l'ipotesi che gli strati hanno tutti la stessa ampiezza, ne consegue che $S^2_{xg} \leqslant (x_{max} - x_{min})^2/(4M^2)$, da cui:

$$
V(\widehat{\mu}_{str};\, prop) \leqslant \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\beta^2_{y/x}\, \frac{(x_{max} - x_{min})^2}{4M^2} + S^2_e\right\}. \qquad (8.24)
$$

La (8.24) mostra un fatto molto importante: al crescere del numero degli strati, il termine

$$
\beta^2_{y/x}\, \frac{(x_{max} - x_{min})^2}{4M^2} \qquad (8.25)
$$

decresce come il quadrato del numero degli strati stessi, e quindi diventa rapidamente piccolo al crescere di M. In genere, basta un valore non elevato di M per rendere il termine (8.25) tanto piccolo da poter essere considerato "trascurabile" rispetto a S^2_e, che invece diviene preponderante nel determinare $V(\widehat{\mu}_{str})$. Ulteriori incrementi del numero di strati, non modificando in alcun

modo il termine S_e^2, riducono di pochissimo $V(\widehat{\mu}_{str})$, e quindi sono inutili agli effetti pratici.

Si può anche dire qualcosa in più sul guadagno che si ha usando il disegno stratificato (proporzionale) rispetto al disegno ssr (quando usato in coppia con la media campionaria). In generale, confrontare $V(\widehat{\mu}_{str};\ prop)$ e $V(\overline{y}_{\mathbf{s}};\ ssr)$ in modo da mettere in evidenza il ruolo svolto dal numero M degli strati è tutt'altro che agevole. Tale confronto è però semplice se si confrontano i valori *massimi* delle varianze di $\widehat{\mu}_{str}$ e di $\overline{y}_{\mathbf{s}}$. Dalle due relazioni $S_y^2 = \beta_{y/x}^2 S_x^2 + S_e^2$ e $S_x^2 \leqslant (x_{max} - x_{min})^2/4$ si deduce infatti che $S_y^2 \leqslant (x_{max} - x_{min})^2/4 + S_e^2$. Usando quindi le approssimazioni

$$S_x^2 \approx \frac{(x_{max} - x_{min})^2}{4}, \quad S_y^2 \approx \frac{(x_{max} - x_{min})^2}{4} + S_e^2$$

si ottiene

$$\begin{aligned} \frac{V(\widehat{\mu}_{str};\ prop)}{V(\overline{y}_{\mathbf{s}};\ ssr)} &\approx \frac{\beta_{y/x}^2 \frac{S_x^2}{M^2} + S_e^2}{S_y^2} \\ &= \frac{\rho_{yx}^2 \frac{S_y^2}{M^2} + S_y^2(1 - \rho_{xy}^2)}{S_y^2} \\ &= \frac{\rho_{yx}^2}{M^2} + (1 - \rho_{xy}^2) \end{aligned} \tag{8.26}$$

essendo ρ_{xy} il coefficiente di correlazione lineare tra $\mathcal{X}$ e $\mathcal{Y}$. La (8.26) è ottenuta, con considerazioni e ipotesi in parte diverse, in Cochran (1977) (pp. 132–134). Ad ogni modo, la relazione (8.26) mette in evidenza che il vantaggio che si ottiene usando il disegno stratificato (proporzionale, basato su un carattere di stratificazione $\mathcal{X}$) rispetto a quello ssr dipende essenzialmente dal termine ρ_{xy}^2/M^2, quantità che decresce rapidamente a 0 al crescere di M. Questo conferma quanto detto in precedenza, ovvero che a meno di casi speciali (valori molto alti di ρ_{xy}^2), il vantaggio di efficienza che si ottiene aumentando il numero di strati diventa rapidamente trascurabile. Questa asserzione è corroborata anche dalle elaborazioni numeriche in Tabella 8.2, in cui sono riportati i valori assunti dalla (8.26) in corrispondenza a differenti valori di ρ_{xy}^2 e M.

Per concludere, un paio di osservazioni. In primo luogo, tra le ipotesi fatte vi è quella che il disegno campionario sia di tipo proporzionale. Risultati simili valgono anche in altri casi, come ad esempio quando si usa l'allocazione di Neyman. In secondo luogo, approssimare le varianze S_{xg}^2, S_x^2 con i loro valori massimi (pari rispettivamente a $(x_{max} - x_{min})^2/(4M^2)$ e $(x_{max} - x_{min})^2/4$) significa usare implicitamente una relazione del tipo $S_{xg}^2 = S_x^2/M^2$. Questo tipo di relazione tra varianze di strato e varianza totale è valida in molti casi, che coinvolgono popolazioni di forma assai differente (Esercizio 8.5). Per approfondimenti si rinvia al lavoro di Cochran (1961).

Tabella 8.2 Valori di $V(\widehat{\mu}_{str};\ prop)/V(\overline{y}_{\mathbf{s}};\ ssr)$ ottenuti dalla (8.26)

$\frac{M}{\rho^2_{xy}}$	*2*	*3*	*3*	*5*	*6*	*10*	*20*	∞
0.30	0.775	0.733	0.719	0.712	0.708	0.703	0.701	0.700
0.40	0.700	0.644	0.625	0.616	0.611	0.604	0.601	0.600
0.50	0.625	0.556	0.531	0.52	0.514	0.505	0.501	0.500
0.75	0.438	0.333	0.297	0.280	0.271	0.258	0.252	0.250
0.9	0.325	0.200	0156	0.136	0.125	0.109	0.102	0.100
0.95	0.288	0.156	0.109	0.08	0.076	0.060	0.052	0.050
0.99	0.258	0.120	0.072	0.050	0.038	0.020	0.012	0.01

8.3 Il problema dell'allocazione nel caso di più caratteri di interesse

Fino ad ora si è sempre supposto che l'interesse della rilevazione statistica sia nella stima della media di un carattere statistico di interesse. Questo, però, è un caso abbastanza raro nelle applicazioni. Molto spesso vi sono k caratteri di interesse, diciamo $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$. Questo complica non poco la notazione. Supponiamo, al solito, che la popolazione sia suddivisa in M strati, rispettivamente di numerosità $N_1, \ldots, N_M$ (e pesi $w_1 = N_1/N, \ldots, w_M = N_M/N$). In generale, indichiamo con y_{jgi} la modalità che il carattere $\mathcal{Y}_j$ ($j = 1, \ldots, k$) assume in corrispondenza dell'unità i (=1, ..., N_g) dello strato g (=1, ..., M). Indichiamo poi con

$$\mu_{jg} = \frac{1}{N_g}\sum_{i=1}^{N_g} y_{jgi}, \quad S^2_{jg} = \frac{1}{N_g - 1}\sum_{i=1}^{N_g}(y_{jgi} - \mu_{jg})^2;$$
$$g = 1, \ldots, M\,;\ j = 1, \ldots, k$$

rispettivamente la media e la varianza corretta del carattere $\mathcal{Y}_j$ nello strato g-mo, e con

$$\mu_j = \sum_{g=1}^{M} w_g\, \mu_{jg}; \quad j = 1, \ldots, k$$

la media di $\mathcal{Y}_j$ nell'intera popolazione.

Se si vogliono stimare le k medie $\mu_1, \ldots, \mu_k$, è naturale considerare i k stimatori

$$\widehat{\mu}_{str,j} = \sum_{g=1}^{M} w_g\, \overline{y}_{jg}$$

dove, con la solita notazione, $\overline{y}_{jg}$ è la media campionaria del carattere $\mathcal{Y}_j$ nello

strato g-mo. Naturalmente, valgono le relazioni

$$E\left[\widehat{\mu}_{str,j}\right] = \mu_j, \;\; V\left(\widehat{\mu}_{str,j}\right) = \sum_{g=1}^{M} \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{jg}^2 \quad j = 1, \ldots, k.$$

Il problema dell'allocazione delle unità campionarie quando si hanno più caratteri è molto più complicato rispetto al caso di un solo carattere. Supponiamo infatti che sia fissata la numerosità campionaria totale n. Se $n_1, \ldots, n_M$ sono scelti in modo da rendere minima la varianza di $\widehat{\mu}_{str,j}$, non è affatto detto che siano automaticamente minimizzate anche le varianze degli altri $k-1$ stimatori $\widehat{\mu}_{str,h}$, con $h \neq j$. Anzi, in generale questo non accade. In sostanza, pertanto, se si devono stimare le medie di k caratteri non esiste un'allocazione delle unità campionarie nei diversi strati che minimizzi simultaneamente le varianze dei k stimatori $\widehat{\mu}_{str,1}, \ldots, \widehat{\mu}_{str,k}$.

In casi come quello appena delineato vi sono diversi modi per uscire dall'*impasse*. Una prima possibilità, che è poi la più frequente sul piano operativo, è quella di identificare il carattere principale di interesse, e nell'allocare le unità campionarie in maniera ottimale rispetto ad esso. In effetti, se è vero che nelle indagini campionarie concrete si rilevano parecchi caratteri di interesse, è anche vero che essi non sono tutti ugualmente interessanti. Il caso più comune è quello in cui vi è un carattere di speciale interesse, che giustifica l'effettuazione di una rilevazione campionaria. È questo il *carattere principale* della rilevazione. Accanto ad esso, vi sono poi altri caratteri che, pur se di interesse, sono in qualche modo "secondari". In casi come questo la soluzione comune è quella di allocare le unità negli strati in modo da minimizzare la varianza dello stimatore $\widehat{\mu}_{str}$ della media del carattere principale. Questo, in realtà, è un principio che vale non solo per il problema dell'allocazione. Sia la scelta della numerosità campionaria totale n che la costruzione degli strati possono essere effettuati facendo esclusivamente riferimento al carattere principale della rilevazione.

Quando in una rilevazione campionaria non vi è un unico carattere principale, il discorso fatto in precedenza viene meno. In questi casi, una soluzione semplice al problema dell'allocazione può essere quella di utilizzare un disegno stratificato proporzionale. Una semplice alternativa, spesso considerata in letteratura (per la verità più sul piano teorico che nelle concrete applicazioni) è quella di far riferimento ad una media ponderata delle varianze degli stimatori $\widehat{\mu}_{str,j}$, $j = 1, \ldots, k$. Sul piano formale, bisogna stabilire k pesi $q_1, \ldots, q_k$, tali che

$$q_j \geqslant 0 \;\; \textit{per ogni } j = 1, \ldots, k; \;\; \sum_{j=1}^{k} q_j = 1$$

e, dato n, determinare $n_1, \ldots, n_M$ in modo da minimizzare la quantità

$$V = \sum_{j=1}^{k} q_j \, V(\widehat{\mu}_{str,j}) = \sum_{j=1}^{k} q_j \left\{ \sum_{g=1}^{M} \left(\frac{1}{n_g} - \frac{1}{N_g}\right) w_g^2 \, S_{jg}^2 \right\}. \qquad (8.27)$$

Invertendo l'ordine di somma, è immediato verificare (Esercizio 8.6) che

$$\sum_{j=1}^{k} q_j \left\{ \sum_{g=1}^{M} \left(\frac{1}{n_g} - \frac{1}{N_g} \right) w_g^2 \, S_{jg}^2 \right\} = \sum_{g=1}^{M} \frac{w_g^2 \, V_g^2}{n_g} - \frac{1}{N} \sum_{g=1}^{M} w_g \, V_g^2 \quad (8.28)$$

con

$$V_g^2 = \sum_{j=1}^{k} q_j \, S_{jg}^2, \quad g = 1, \ldots, M. \quad (8.29)$$

Il problema di minimizzare la (8.27), con $n_1 + \cdots + n_M = n$, si può riformulare quindi nel modo seguente:

$$\begin{cases} \textit{minimizzare} : \sum_{g=1}^{M} \frac{w_g^2 V_g^2}{n_g} \\ \textit{con il vincolo} : \sum_{g=1}^{M} n_g = n \end{cases} . \quad (8.30)$$

Usando esattamente la stessa tecnica della Sezione 7.4.1, è immediato verificare (Esercizio 8.7) che la soluzione del problema (8.30) è del tipo:

$$n_g = n \, \frac{w_g \, V_g}{\sum_{h=1}^{M} w_h \, V_h}; \quad g = 1, \ldots, M. \quad (8.31)$$

Il punto più debole di questo approccio è nel modo in cui i pesi andrebbero scelti. In linea di principio, il peso q_j rappresenta l'importanza relativa del carattere $\mathcal{Y}_j$. In questo modo, però, la soluzione (8.31) dipende dalle unità di misura usate per i diversi caratteri. Ad es., misurare le stature in centimetri anziché in metri moltiplica la corrispondente varianza per un fattore 10000, e quindi modifica i valori delle numerosità campionarie ottimali (8.31). Un possibile modo per ovviare a questo difetto consiste nel rendere i pesi inversamente proporzionali alle deviazioni *standard* dei corrispondenti caratteri. In questo modo, però, viene meno il significato stesso dei pesi. Per ulteriori approfondimenti su questo punto si rinvia al volume Cicchitelli *e altri* (1992).

8.4 Stimatori di tipo quoziente nel campionamento stratificato

Fino ad ora in coppia con il disegno stratificato è stato sempre utilizzato lo stimatore $\widehat{\mu}_{str}$, di cui sono state studiate le proprietà. Vi sono però circostanze in cui può essere vantaggioso usare stimatori di tipo differente, che incorporano informazioni ausiliarie. Il caso più semplice è quello in cui tali informazioni siano rappresentate da un carattere $\mathcal{X}$ le cui modalità siano note *a priori* su tutte le unità della popolazione. Per evitare confusioni con la notazione usata nei paragrafi precedenti, sottolineiamo subito che il carattere ausiliario $\mathcal{X}$ a cui qui si fa riferimento *non* è lo stesso usato come carattere di stratificazione.

Per quanto riguarda il carattere di interesse $\mathcal{Y}$, la notazione è esattamente quella usata fino ad ora. Per quanto attiene al carattere ausiliario $\mathcal{X}$, sia x_{gi} la modalità che esso assume in corrispondenza dell'unità i $(= 1, \ldots, N_g)$ dello strato g $(= 1, \ldots, M)$. Indichiamo poi con

$$\mu_{xg} = \frac{1}{N_g} \sum_{i=1}^{N_g} x_{gi}, \quad S^2_{xg} = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (x_{gi} - \mu_{xg})^2,$$

$$S_{xy\,g} = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (x_{gi} - \mu_{xg})\,(y_{gi} - \mu_{yg})$$

rispettivamente la media, la varianza corretta di $\mathcal{X}$ e la covarianza corretta tra $\mathcal{X}$ e $\mathcal{Y}$ nello strato g-mo $(g = 1, \ldots, M)$.

Se dalla popolazione si seleziona un campione stratificato $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$, in cui il sottocampione $\mathbf{s}_g$ dello strato g-mo ha numerosità n_g, sia infine

$$\overline{x}_g = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} x_{gi}$$

la media campionaria di $\mathcal{X}$ nello strato g-mo.

8.4.1 Stimatore per quoziente separato

Lo stimatore $\widehat{\mu}_{str}$ si fonda su un approccio molto semplice:

- stimare le medie $\mu_{y1}, \ldots, \mu_{yM}$ di $\mathcal{Y}$ separatamente strato per strato, tramite le corrispondenti medie campionarie $\overline{y}_1, \ldots, \overline{y}_M$;
- ricombinare le M stime in (i) per ottenere una stima di μ_y.

Alla base dello stimatore per quoziente separato vi è un'idea elementare: cercare di usare il carattere ausiliario $\mathcal{X}$ per migliorare le stime di $\mu_{y1}, \ldots, \mu_{yM}$ che si ottengono tramite le medie campionarie $\overline{y}_1, \ldots, \overline{y}_M$. Se tra i due caratteri $\mathcal{Y}$ e $\mathcal{X}$ intercorre, in via approssimata, una relazione di "quasi-proporzionalità" all'interno di ciascuno strato, si può pensare di stimare μ_{yg} con uno stimatore per quoziente (ristretto allo strato g-mo) del tipo

$$\widehat{\mu}_{q\,g} = \frac{\overline{y}_g}{\overline{x}_g}\,\mu_{xg}, \quad g = 1, \ldots, M. \tag{8.32}$$

Gli M stimatori (8.32) vanno poi ricombinati assieme, per produrre una stima della media μ_y della popolazione. Per effettuare tale ricombinazione, seguiamo esattamente lo stesso approccio che porta allo stimatore $\widehat{\mu}_{str}$. L'unica differenza è che ora si hanno gli M stimatori $\widehat{\mu}_{q\,1}, \ldots, \widehat{\mu}_{q\,M}$ anziché le medie campionarie di strato $\overline{y}_1, \ldots, \overline{y}_M$. Si ha in questo modo lo *stimatore per quoziente separato*, che assume la forma:

$$\widehat{\mu}_{q\,sep} = \sum_{g=1}^{M} w_g\,\widehat{\mu}_{q\,g}. \tag{8.33}$$

Le proprietà dello stimatore per quoziente separato (8.33) si ottengono facilmente a partire da quelle dello stimatore quoziente visto nel Capitolo 6. Poniamo:

$$R_g = \frac{\mu_{yg}}{\mu_{xg}},\ \widehat{R}_g = \frac{\overline{y}_g}{\overline{x}_g};\ \ g = 1, \ldots, M$$

così che si può scrivere

$$\widehat{\mu}_{q\,sep} = \sum_{g=1}^{M} w_g\, \widehat{R}_g\, \mu_{xg}.$$

Essendo $E[\widehat{R}_g] \neq R_g$, si ha che $E[\widehat{\mu}_{q\,g}] \neq \mu_{yg}$, e quindi lo stimatore (8.33) è distorto. La sua varianza esatta, inoltre, non è esprimibile in forma esplicita (tranne casi eccezionali). Come conseguenza di quanto visto nel Capitolo 6, se le numerosità campionarie di strato $n_1, \ldots, n_M$ sono sufficientemente grandi si può scrivere

$$\begin{aligned}
E[\widehat{\mu}_{q\,g}] &= E[\widehat{R}_g]\, \mu_{xg} \approx R_g\, \mu_{xg} = \mu_{yg};\\
V(\widehat{\mu}_{q\,g}) &= V(\widehat{R}_g)\, \mu_{xg}^2 \approx \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left\{S_{yg}^2 + R_g^2\, S_{xg}^2 - 2\, R_g\, S_{xy\,g}\right\};\\
MSE(\widehat{\mu}_{q\,g}) &= MSE(\widehat{R}_g)\, \mu_{xg}^2 \approx \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left\{S_{yg}^2 + R_g^2\, S_{xg}^2 - 2\, R_g\, S_{xy\,g}\right\}.
\end{aligned}$$

De queste relazioni si ottiene facilmente la seguente proposizione.

Proposizione 8.2. *Se il disegno campionario è stratificato valgono le seguenti relazioni:*

$$E[\widehat{\mu}_{q\,sep}] = \sum_{g=1}^{M} w_g\, E[\widehat{\mu}_{q\,g}] \approx \sum_{g=1}^{M} w_g\, \mu_{yg} = \mu_y;$$

$$\begin{aligned}
V(\widehat{\mu}_{q\,sep}) &= \sum_{g=1}^{M} w_g^2\, V(\widehat{\mu}_{q\,g})\\
&\approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left\{S_{yg}^2 + R_g^2\, S_{xg}^2 - 2\, R_g\, S_{xy\,g}\right\};
\end{aligned}$$

$$\begin{aligned}
MSE(\widehat{\mu}_{q\,sep}) &\approx V(\widehat{\mu}_{q\,sep})\\
&\approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left\{S_{yg}^2 + R_g^2\, S_{xg}^2 - 2\, R_g\, S_{xy\,g}\right\}.
\end{aligned}$$

Sulla base dei risultati del Capitolo 6 non è difficile costruire uno stimatore della varianza (e quindi dell'errore quadratico medio) dello stimatore quoziente separato. Infatti, procedendo come nella Sezione 6.3 è facile verificare

che:

$$S_{yg}^2 + R_g^2 S_{xg}^2 - 2 R_g S_{xy\,g} = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} + R_g x_{gi})^2$$

per cui si può scrivere:

$$V(\widehat{\mu}_{q\,g}) \approx \left(\frac{1}{n_g} - \frac{1}{N_g} \right) \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} + R_g x_{gi})^2.$$

Procedendo lungo le linee della Sezione 6.3, come "ragionevole" stimatore di $V(\widehat{\mu}_{q\,g})$ si può fare riferimento a

$$\widehat{V}(\widehat{\mu}_{q\,g}) = \left(\frac{1}{n_g} - \frac{1}{N_g} \right) \frac{1}{n_g - 1} \sum_{i \in \mathbf{s}_g} (y_{gi} - \widehat{R}_g x_{gi})^2.$$

In definitiva, quindi, come stimatore della varianza (ed anche dell'errore quadratico medio) di $\widehat{\mu}_{q\,g}$ avremo il seguente:

$$\begin{aligned} \widehat{V}(\widehat{\mu}_{q\,g}) &= \sum_{g=1}^{M} w_g^2 \, \widehat{V}(\widehat{\mu}_{q\,g}) \\ &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) \left\{ \frac{1}{n_g - 1} \sum_{i \in \mathbf{s}_g} (y_{gi} - \widehat{R}_g x_{gi})^2 \right\}. \end{aligned} \quad (8.34)$$

Utilizzando poi l'approssimazione normale per la distribuzione di probabiltà dello stimatore per quoziente separato, sulla base della (8.34) è facile costruire un intervallo di confidenza approssimato per μ_y.

Prima di concludere, vale la pena ritornare brevemente sugli aspetti logici riguardanti l'appropriatezza dello stimatore per quoziente separato. Si è già detto che esso fornisce risultati buoni se tra i due caratteri $\mathcal{Y}$ e $\mathcal{X}$ intercorre una relazione approssimata di "quasi-proporzionalità" all'interno di ciascuno strato. Ciò è grosso modo equivalente a richiedere che, all'interno di ogni strato, la retta di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ passi approssimativamente per l'origine. In questo caso, come visto nel Capitolo 6, il coefficiente di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ nello strato g-mo è proprio uguale a $R_g = \mu_{yg}/\mu_{xg}$. Dal momento che le quantità R_g sono stimate *separatamente in ogni strato* (con le $\widehat{R}_g$), lo stimatore $\widehat{\mu}_{q\,sep}$ non richiede nessuna ipotesi aggiuntiva. Le quantità R_g possono essere uguali o differenti, ma ciò ha scarsa rilevanza sulla qualità dello stimatore $\widehat{\mu}_{q\,sep}$, dal momento che essi sono stimati separatamente in ogni strato.

8.4.2 Stimatore per quoziente combinato

Lo stimatore per quoziente separato (8.33) si basa essenzialmente sull'uso dello stimatore per quoziente a livello di singoli strati. Un'alternativa abbastanza

semplice è quella di usare le idee di base del metodo del quoziente a livello dello stimatore $\widehat{\mu}_{str}$. Come più volte sottolineato nel Capitolo 6, lo stimatore per quoziente $\widehat{\mu}_q$ nel caso di campionamento ssr ha una struttura del tipo

$$\widehat{\mu}_q = \frac{\overline{y}_{\mathbf{s}}}{\overline{x}_{\mathbf{s}}}\,\mu_x = \left(stima\ di\ \frac{\mu_y}{\mu_x}\right) \times \mu_x. \tag{8.35}$$

Il fatto che nella (8.35) compaiono le medie campionarie $\overline{y}_{\mathbf{s}}$, $\overline{x}_{\mathbf{s}}$ dipende solo dal fatto che, essendo il disegno campionario di tipo ssr, esse sono usate come stimatori "naturali" rispettivamente di μ_y e μ_x.

Se il disegno campionario è di tipo stratificato, come stimatori "naturali" di μ_y e μ_x si possono usare rispettivamente:

$$\widehat{\mu}_{str\,y} = \sum_{g=1}^{M} w_g\,\overline{y}_g, \quad \widehat{\mu}_{str\,x} = \sum_{g=1}^{M} w_g\,\overline{x}_g$$

per cui come stimatore di $R = \mu_y/\mu_x$ si farà riferimento a:

$$\widehat{R}_{str} = \frac{\widehat{\mu}_{str\,y}}{\widehat{\mu}_{str\,x}}. \tag{8.36}$$

Usando sempre la struttura di base (8.35) si ottiene lo *stimatore per quoziente combinato*:

$$\begin{aligned} \widehat{\mu}_{q\,com} &= \widehat{R}_{str}\,\mu_x \\ &= \frac{\sum_{g=1}^{M} w_g\,\overline{y}_g}{\sum_{g=1}^{M} w_g\,\overline{x}_g}\,\mu_x. \end{aligned} \tag{8.37}$$

Lo studio delle proprietà esatte dello stimatore (8.37) è complicato, in quanto l'avere al denominatore lo stimatore $\widehat{\mu}_{str\,x}$, che varia al variare dei dati campionari, fa sì che il valore atteso di $\widehat{\mu}_{q\,com}$ non sia (esclusi alcuni casi eccezionali) esprimibile in forma esplicita. Alcuni risultati si possono ottenere usando lo stesso approccio del Capitolo 6. In particolare, usando la notazione in (8.36), è facile vedere (Esercizio 8.8) che la distorsione di $\widehat{\mu}_{q\,com}$ assume la seguente espressione:

$$B(\widehat{\mu}_{q\,com}) = -C(\widehat{R}_{str},\,\widehat{\mu}_{str\,x}). \tag{8.38}$$

Per numerosità campionarie totali n abbastanza "grandi", usando sostanzialmente gli stessi argomenti già adoperati sia nel Capitolo 6 che per lo stimatore quoziente separato si ha che $E[\widehat{R}_{str}]$ è approssimativamente uguale a $R = \mu_y/\mu_x$. Si può quindi enunciare la seguente proposizione.

Proposizione 8.3. *Se il disegno campionario è stratificato valgono le seguenti relazioni:*

$$E[\widehat{\mu}_{q\,com}] = E\left[\widehat{R}_{str}\right]\,\mu_x \approx R\,\mu_x = \mu_y; \tag{8.39}$$

$$V(\widehat{\mu}_{q\,com}) \approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right)\left\{S_{yg}^2 + R^2\,S_{xg}^2 - 2\,R\,S_{xy\,g}\right\}; \tag{8.40}$$

$$MSE(\widehat{\mu}_{q\,sep}) \approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right)\left\{S_{yg}^2 + R^2\,S_{xg}^2 - 2\,R\,S_{xy\,g}\right\}. \tag{8.41}$$

Dimostrazione. Gli argomenti sono gli stessi usati per il disegno ssr. Si può anzitutto scrivere

$$\begin{aligned}\widehat{R}_{str} - R &= \frac{\widehat{\mu}_{str\,y} - R\,\widehat{\mu}_{str\,x}}{\widehat{\mu}_{str\,x}} \\ &= \frac{\widehat{\mu}_{str\,y} - R\,\widehat{\mu}_{str\,x}}{\mu_x} + (\widehat{\mu}_{str\,y} - R\,\widehat{\mu}_{str\,x})\left(\frac{1}{\widehat{\mu}_{str\,x}} - \frac{1}{\mu_x}\right).\end{aligned}$$

Con uno sviluppo di Taylor nel punto μ_x si ottiene poi

$$\frac{1}{\widehat{\mu}_{str\,x}} = \frac{1}{\mu_x} + Resto \approx \frac{1}{\mu_x}$$

da cui segue, in definitiva, che

$$\widehat{R}_{str} \approx R + \frac{\widehat{\mu}_{str\,y} - R\,\widehat{\mu}_{str\,x}}{\mu_x}$$

e quindi

$$\widehat{\mu}_{q\,com} \approx \mu_y + \widehat{\mu}_{str\,y} - R\,\widehat{\mu}_{str\,x}. \tag{8.42}$$

Dalla (8.42) si ha subito:

$$\begin{aligned}E\,[\widehat{\mu}_{q\,com}] &\approx \mu_y + E\,[\widehat{\mu}_{str\,y}] - R\,E\,[\widehat{\mu}_{str\,x}] \\ &= \mu_y + (\mu_y - R\,\mu_x) \\ &= \mu_y\end{aligned}$$

ossia la (8.39).

Per quanto concerne la (8.40), è sufficiente osservare che, sempre per la (8.42), si può scrivere

$$\widehat{\mu}_{q\,com} \approx \mu_y + \sum_{g=1}^{M} w_g\,(\overline{y}_g - R\,\overline{x}_g)$$

da cui si ottiene

$$\begin{aligned} V(\widehat{\mu}_{q\,com}) &\approx V\left(\sum_{g=1}^{M} w_g\,(\overline{y}_g - R\,\overline{x}_g)\right) \\ &= \sum_{g=1}^{M} w_g^2\,V(\overline{y}_g - R\,\overline{x}_g) \\ &= \sum_{g=1}^{M} w_g^2\,\{V(\overline{y}_g) + R^2\,V(\overline{x}_g) - 2\,R\,C(V(\overline{y}_g,\,\overline{y}_g)\} \\ &= \sum_{g=1}^{M} w_g^2\,\left(\frac{1}{n_g} - \frac{1}{N_g}\right)\,\{S_{yg}^2 + R^2\,S_{xg}^2 - 2\,R\,S_{xy\,g}\}. \end{aligned}$$

Infine, la (8.41) è una immediata conseguenza di (8.39) e (8.40). □

La (8.40) suggerisce anche il seguente stimatore della varianza di $\widehat{\mu}_{q\,com}$:

$$\widehat{V}(\widehat{\mu}_{q\,com}) = \sum_{g=1}^{M} w_g^2\,\left(\frac{1}{n_g} - \frac{1}{N_g}\right)\,\left\{\widehat{s}_{yg}^2 + \widehat{R}_{str}^2\,\widehat{s}_{xg}^2 - 2\,\widehat{R}_{str}\,\widehat{s}_{xy\,g}\right\}$$

dove $\widehat{R}_{str}$ è dato dalla (8.36), e

$$\widehat{s}_{yg}^2 = \frac{1}{n_g - 1}\sum_{i\in\mathbf{s}_g}(y_{gi} - \overline{y}_g)^2,\quad \widehat{s}_{xg}^2 = \frac{1}{n_g - 1}\sum_{i\in\mathbf{s}_g}(x_{gi} - \overline{x}_g)^2, \quad (8.43)$$

$$\widehat{s}_{xy\,g} = \frac{1}{n_g - 1}\sum_{i\in\mathbf{s}_g}(x_{gi} - \overline{x}_g)\,(y_{gi} - \overline{y}_g), \quad (8.44)$$

(con $g = 1, \ldots, M$) sono rispettivamente la varianze campionarie corrette di $\mathcal{Y}$ e $\mathcal{X}$ e la covarianza campionaria corretta tra $\mathcal{Y}$ e $\mathcal{X}$, nello strato g-mo.

Per quanto attiene all'efficienza dei due stimatori per quoziente separato e combinato, essa dipende da due fattori: i rapporti $R_g = \mu_{yg}/\mu_{xg}$ tra le medie di strato, e le numerosità campionarie di strato n_g. Lo stimatore per quoziente separato, in generale, fornisce risultati tanto migliori quanto più i rapporti R_g assumono valori "simili". Se si ha ragione di ritenere che i rapporti R_g assumono valori di molto differenti nei diversi strati, lo stimatore $\widehat{\mu}_{q\,com}$ avrà un'efficienza anche parecchio inferiore a $\widehat{\mu}_{q\,sep}$. D'altra parte, non è da trascurare il ruolo delle numerosità campionarie n_g. Se queste sono piccole, i rapporti campionari $\widehat{R}_g = \overline{y}_g/\overline{x}_g$ forniranno in genere stime piuttosto imprecise delle R_g, così che l'efficienza dello stimatore per quoziente separato tende a degradarsi. In tali condizioni sarà quindi preferibile usare lo stimatore per quoziente combinato.

8.5 Stimatori per regressione nel campionamento stratificato

In maniera del tutto simile a quanto fatto nel precedente paragrafo, si possono introdurre gli stimatori per regressione separato e combinato. Supponiamo sempre che vi sia un carattere $\mathcal{X}$ le cui modalità siano note *a priori* su tutte le unità della popolazione. La notazione che useremo qui è esattamente la stessa usata nel paragrafo precedente. Inoltre, indicheremo con:

$$b_{y/x\,g} = \frac{S_{xy\,g}}{S^2_{xg}}$$

il coefficiente di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ nello strato g-mo ($g = 1, \ldots, M$).

Dalla popolazione viene selezionato un campione stratificato $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$, in cui il sottocampione $\mathbf{s}_g$ dello strato g-mo ha numerosità n_g. In aggiunta alle grandezze introdotte nel paragrafo precedente (in particolare, le (8.43), (8.44)) indichiamo con

$$\widehat{b}_{y/x\,g} = \frac{\widehat{s}_{xy\,g}}{\widehat{s}^2_{xg}}\,, \quad g = 1, \ldots, M$$

il coefficiente di regressione campionario di $\mathcal{Y}$ rispetto a $\mathcal{X}$ nello strato g ($= 1, \ldots, M$).

8.5.1 Stimatore per regressione separato

Lo stimatore per regressione separato, che d'ora in avanti indicheremo con $\widehat{\mu}_{reg\,sep}$ si fonda su un approccio del tutto simile a quello che porta, nel caso di disegno ssr, allo stimatore di regressione "usuale". Infatti, nel caso di disegno campionario stratificato, un'estensione del tutto naturale dello stimatore alle differenze introdotto nel Capitolo 5 è il seguente:

$$\widehat{\mu}_{gd\,sep} = \sum_{g=1}^{M} w_g \left\{\overline{y}_g - c_g \left(\overline{x}_g - \mu_{xg}\right)\right\} \tag{8.45}$$

dove $c_1, \ldots, c_M$ sono arbitrari numeri reali. Non è difficile verificare (Esercizio 8.9) che $\widehat{\mu}_{gd\,sep}$ è uno stimatore corretto di μ_y, e che la sua varianza è pari a

$$V(\widehat{\mu}_{gd\,sep}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \{S^2_{yg} + c_g^2\, S^2_{xg} - 2\, c_g\, S_{xy\,g}\}. \tag{8.46}$$

Un criterio molto naturale per scegliere $c_1, \ldots, c_M$ consiste nell'assegnare ad essi i valori che rendono minima la (8.46), in modo da avere la massima

efficienza di stima. È immediato verificare (Esercizio 8.10) che tali valori ottimali sono i coefficienti di regressione di $\mathcal{Y}$ rispetto a $\mathcal{X}$ nei diversi strati. In simboli:

$$c_1 = b_{y/x\,1},\ \ldots,\ c_M = b_{y/x\,M}.$$

I coefficienti $b_{y/x\,g}$ sono incogniti, per cui questa strada non è realmente percorribile. Tuttavia, si può pensare di stimarli su base campionaria, tramite i corrispondenti coefficienti di regressione campionaria $\widehat{b}_{y/x\,g}$. Si ottiene in questo modo lo *stimatore per regressione separato*, che assume la forma:

$$\begin{aligned}\widehat{\mu}_{reg\,sep} &= \sum_{g=1}^{M} w_g \left\{\overline{y}_g - \widehat{b}_{y/x\,g}\,(\overline{y}_g - \mu_{yg})\right\} \\ &= \sum_{g=1}^{M} w_g\,\widehat{\mu}_{reg\,g} \qquad (8.47)\end{aligned}$$

in cui

$$\widehat{\mu}_{reg\,g} = \overline{y}_g - \widehat{b}_{y/x\,g}\,(\overline{y}_g - \mu_{yg}),\ \ g = 1,\ \ldots,\ M$$

è lo stimatore per regressione "usuale" della media μ_{yg} dello strato g-mo. La (8.47) mostra che lo stimatore per regressione separato è null'altro che una media (ponderata con pesi w_g) degli stimatori per regressione delle medie dei diversi strati in cui è suddivisa la popolazione.

Le proprietà dello stimatore per regressione separato (8.47) si ottengono in modo simile a quanto visto per lo stimatore per quoziente separato. Sia

$$\rho_{xy\,g} = \frac{S_{xy\,g}}{S_{xg}\,S_{yg}}$$

il coefficiente di correlazione lineare tra $\mathcal{X}$ e $\mathcal{Y}$ nello strato g. Come conseguenza di quanto visto nel Capitolo 5 (e in particolare della Proposizione 5.2) si ha che

$$\begin{aligned}E[\widehat{\mu}_{reg\,g}] &\approx \mu_{yg}; \\ V(\widehat{\mu}_{reg\,g}) &\approx \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2(1 - \rho_{xy\,g}^2); \\ MSE(\widehat{\mu}_{reg\,g}) &\approx \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2(1 - \rho_{xy\,g}^2).\end{aligned}$$

Pertanto, è facile provare la seguente proposizione.

Proposizione 8.4. *Se il disegno campionario è stratificato valgono le seguenti relazioni:*

$$E[\widehat{\mu}_{reg\,sep}] = \sum_{g=1}^{M} w_g\, E[\widehat{\mu}_{reg\,g}] \approx \sum_{g=1}^{M} w_g\, \mu_{yg} = \mu_y;$$

$$V(\widehat{\mu}_{reg\,sep}) = \sum_{g=1}^{M} w_g^2\, V(\widehat{\mu}_{reg\,g}) \approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 (1 - \rho_{xy\,g}^2);$$

$$MSE(\widehat{\mu}_{reg\,sep}) \approx V(\widehat{\mu}_{reg\,sep}) \approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 (1 - \rho_{xy\,g}^2).$$

Non è neanche difficile costruire uno stimatore della varianza (e quindi dell'errore quadratico medio) dello stimatore per regressione separato. Procedendo infatti come nella Sezione 5.5, come stimatore di $V(\widehat{\mu}_{reg\,sep})$ si può usare:

$$\widehat{V}(\widehat{\mu}_{reg\,g}) = \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(\widehat{s}_{yg}^2 - \widehat{b}_{y/x\,g}^2\, \widehat{s}_{xg}^2\right).$$

Utilizzando la Proposizione 8.4, come stimatore di $V(\widehat{\mu}_{reg\,sep})$ (e di $MSE(\widehat{\mu}_{reg\,sep})$) si ha il seguente:

$$\widehat{V}(\widehat{\mu}_{reg\,g}) = \sum_{g=1}^{M} w_g^2\, \widehat{V}(\widehat{\mu}_{reg\,g}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(\widehat{s}_{yg}^2 - \widehat{b}_{y/x\,g}^2 \widehat{s}_{xg}^2\right).$$

8.5.2 Stimatore per regressione combinato

Come visto nel Capitolo 5, lo stimatore di regressione è stato costruito a partire dallo stimatore alle differenze $\widehat{\mu}_{d,\,c} = \overline{y}_{\mathbf{s}} - c(\overline{x}_{\mathbf{s}} - \mu_x)$, determinando dapprima il valore di c che minimizza la varianza di $\widehat{\mu}_{d,\,c}$, e poi stimando tale quantità sulla base dei dati campionari. Lo stimatore di regressione combinato può essere costruito a partire da considerazioni del tutto simili. Il punto di partenza è costituito dal seguente stimatore di μ_y:

$$\widehat{\mu}_{gd\,c} = \widehat{\mu}_{str\,y} - c\,(\widehat{\mu}_{str\,y} - \mu_x) = \sum_{g=1}^{M} w_g\,(\overline{y}_g - c\,(\overline{x}_g - \mu_x)) \qquad (8.48)$$

essendo c una arbitraria costante reale. Lo stimatore (8.48) può essere visto come un caso speciale dello stimatore (8.45), in cui $c_1, \ldots, c_M$ sono tutti uguali. È facile provare (Esercizio 8.11) che lo stimatore $\widehat{\mu}_{gd\,c}$ è corretto, e che la sua varianza è pari a

$$V(\widehat{\mu}_{gd\,c}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(S_{yg}^2 + c^2\, S_{xg}^2 - 2\,c\, S_{xy\,g}\right). \qquad (8.49)$$

Per quanto riguarda la costante c, una scelta molto naturale consiste nello scegliere il valore che rende minima la varianza dello stimatore (8.48), ovvero che ne massimizza l'efficienza. Indichiamo con b_o tale valore. È immediato verificare (Esercizio 8.12) che esso risulta pari a

$$b_o = \frac{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xy\,g}}{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xg}^2}. \tag{8.50}$$

Il valore effettivamente assunto dalla (8.50) è incognito, in quanto dipende dalle incognite quantità $S_{xy\,g}$ (oltre che da S_{xg}^2). Sostituendo alle $S_{xy\,g}$ e S_{xg}^2 le loro controparti campionarie, pari rispettivamente a $\widehat{s}_{xy\,g}$, $\widehat{s}_{xg}^2$, si ottiene la seguente stima campionaria di b_o:

$$\widehat{b}_o = \frac{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \widehat{s}_{xy\,g}}{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \widehat{s}_{xg}^2}. \tag{8.51}$$

Sostituendo infine nella (8.48) a c il valore $\widehat{b}_o$ dato dalla (8.51) si ottiene lo *stimatore per regressione combinato*:

$$\widehat{\mu}_{reg\,com} = \sum_{g=1}^{M} w_g \left(\overline{y}_g - \widehat{b}_o \left(\overline{x}_g - \mu_x\right)\right). \tag{8.52}$$

Lo stimatore (8.52) è distorto, e lo studio delle sue proprietà esatte assai complicato. Tuttavia, un suo studio approssimato non presenta difficoltà di rilievo. Usando la stessa tecnica della Proposizione 8.3, se la numerosità campionaria n è sufficientemente elevata, si può scrivere in via approssimata $\widehat{b}_o \approx b_o$, e quindi anche

$$\widehat{\mu}_{reg\,com} \approx \sum_{g=1}^{M} w_g \left(\overline{y}_g - b_o \left(\overline{x}_g - \mu_x\right)\right).$$

Usando in buona sostanza gli stessi argomenti già adoperati per lo stimatore quoziente combinato, si ottiene quindi la seguente proposizione.

Proposizione 8.5. *Se il disegno campionario è stratificato valgono le seguenti relazioni:*

$$E[\widehat{\mu}_{reg\,com}] \approx E\left[\sum_{g=1}^{M} w_g \left(\overline{y}_g - b_o \left(\overline{x}_g - \mu_x\right)\right)\right] = \mu_y;$$

$$\begin{aligned} V(\widehat{\mu}_{reg\,com}) &\approx V\left(\sum_{g=1}^{M} w_g \left(\overline{y}_g - b_o \left(\overline{x}_g - \mu_x\right)\right)\right) \\ &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(S_{yg}^2 + b_o^2\, S_{xg}^2 - 2\, b_o\, S_{xy\,g}\right); \end{aligned} \tag{8.53}$$

$$MSE(\widehat{\mu}_{reg\,com}) \approx V(\widehat{\mu}_{reg\,com}) \approx \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(S_{yg}^2 + b_o^2\, S_{xg}^2 - 2\, b_o\, S_{xy\,g}\right).$$

L'espressione approssimata (8.53) suggerisce infine il seguente stimatore di $V(\widehat{\mu}_{reg\,com})$:

$$\widehat{V}(\widehat{\mu}_{reg\,com}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) \left(\widehat{s}_{yg}^2 + \widehat{b}_o^2\, \widehat{s}_{xg}^2 - 2\, \widehat{b}_o\, \widehat{s}_{xy\,g}\right).$$

Per quanto riguarda l'efficienza dei due stimatori per regressione separato e combinato, iniziamo con l'osservare che, posto

$$q_g = \frac{w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xg}^2}{\sum_{h=1}^{M} w_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{xh}^2};\ g = 1, \ldots, M$$

e tenendo conto che $b_{y/x\,g} = S_{xy\,g}/S_{xg}^2$, la (8.50) si può scrivere come

$$b_o = \sum_{g=1}^{M} q_g\, b_{y/x\,g}.$$

In altri termini, b_o è una media dei coefficienti di regressione nei diversi strati, ponderati con i pesi q_g. Tenendo poi conto che

$$b_{y/x\,g}^2\, S_{xg}^2 = b_{y/x\,g}\, S_{xy\,g},\ \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xy\,g} = b_o \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xg}^2$$

dalle Proposizioni 8.4, 8.5 si desume subito la seguente relazione:

$$\begin{aligned} V(\widehat{\mu}_{reg\,com}) - V(\widehat{\mu}_{reg\,sep}) &\approx \left\{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xg}^2\right\} \\ &\quad \times \left\{\sum_{g=1}^{M} q_g\, b_{y/x\,g}^2 - b_o^2\right\} \\ &= \left\{\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{xg}^2\right\} \\ &\quad \times \sum_{g=1}^{M} q_g\, (b_{y/x\,g} - b_o)^2. \end{aligned} \tag{8.54}$$

La (8.54) mostra che la differenza $V(\widehat{\mu}_{reg\,com}) - V(\widehat{\mu}_{reg\,sep})$ è sempre (in via approssimata) non negativa, e quindi che $V(\widehat{\mu}_{reg\,sep})$ è più piccola di $V(\widehat{\mu}_{reg\,com})$. Lo stimatore per regressione separato è quindi più efficiente di quello combinato. Quest'affermazione va presa *cum grano salis*, per una fondamentale ragione: le espressioni usate per le varianze degli stimatori $\widehat{\mu}_{reg\,com}$ e $\widehat{\mu}_{reg\,sep}$ sono valide solo in via approssimata. In generale, quindi, se le numerosità di strato sono abbastanza piccole, l'approssimazione usata per $V(\widehat{\mu}_{reg\,sep})$ può essere anche molto cattiva, e quindi la conclusione a cui si è giunti è del tutto fuorviante. Come linea guida si può affermare che se le numerosità campionarie di strato sono piccole, o se i coefficienti $b_{y/x\,g}$ assumono valori simili nei diversi strati, sarà preferibile usare lo stimatore per regressione combinato. Se invece i coefficienti $b_{y/x\,g}$ sono molto diversi tra loro, a meno che le numerosità campionarie di strato siano piccole, è preferibile usare lo stimatore per regressione separato.

8.6 Post-Stratificazione

8.6.1 Aspetti di base

Per poter usare un disegno di tipo stratificato è necessario conoscere a priori quali unità formino gli strati in cui è suddivisa la popolazione (pre-stratificazione). La post-stratificazione viene impiegata quando questa informazione non è disponibile, ma si conosce solo *quante* unità formano ciascuno strato. Per sapere a quale strato appartenga un'unità è necessario osservare l'unità stessa.

Formalmente, ciò che è noto sono solo le numerosità N_1, N_2, ..., N_M dei diversi strati. Non sono invece noti i valori della variabile di stratificazione usata (o delle variabili di stratificazione, se ne sono usate più di una). Questi ultimi sono *osservabili* solo sulle unità del campione. Pertanto, solo dopo aver osservato un'unità campionaria è possibile conoscere il relativo strato di appartenenza.

In tali condizioni la soluzione più immediata consiste nel ricorrere al campionamento casuale semplice, e nell'utilizzare successivamente, a livello di stima, l'informazione relativa al carattere di stratificazione. Dopo aver estratto il campione, oltre alla variabile di interesse $\mathcal{Y}$ si osserva anche lo strato a cui appartiene ciascuna delle unità del campione. Tale stratificazione *a posteriori* viene utilizzata come strumento ausiliario indiretto per l'inferenza.

Sia $\mathbf{s}$ il campione di ampiezza n selezionato dalla popolazione secondo un disegno ssr, e, come detto, supponiamo di conoscere, da fonti censuarie o amministrative, la dimensione N_g degli strati della popolazione. Il campione $\mathbf{s}$ può essere suddiviso in M sottocampioni (post-strati campionari) $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$. Il generico sottocampione s_g ($g = 1, \ldots, M$) di numerosità n_g è formato da tutte le unità campionarie che appartengono allo strato

g. Denotiamo inoltre con $\mathbf{n} = (n_1, \ldots, n_M)$ la M-pla delle numerosità dei sottocampioni. Chiaramente, è $n_1 + \cdots + n_M = n$.

Si osservi che essendo le numerosità dei sottocampioni note solo dopo l'estrazione del campione, la M-pla $\mathbf{n} = (n_1, \ldots, n_M)$ è la determinazione di una variabile aleatoria.

Supponiamo, per il momento, che il campione $\mathbf{s}$ contenga almeno un'unità di ciascuno strato, così che i numeri n_1, ..., n_M sono tutti positivi. Lo *stimatore post-stratificato* della media della popolazione μ_y è dato da

$$\widehat{\mu}_{ps} = \sum_{g=1}^{M} w_g \overline{y}_g \tag{8.55}$$

dove $\overline{y}_g$ e $w_g = N_g/N$ rappresentano rispettivamente la media campionaria e il peso dello strato g-esimo.

Esempio 8.3. Si consideri ancora l'Esempio 7.1, in cui si fa riferimento alla popolazione di $N = 1570$ studenti universitari del *file* `stature.txt`; nel *file* stesso sono riportati numero di matricola, sesso e statura di ciascun studente. Convenzionalmente lo strato 1 è quello degli studenti maschi, e lo strato 2 quello degli studenti femmine. Supponiamo di non sapere, *a priori*, quali studenti siano maschi e quali femmine, ma solo che nella popolazione vi sono $N_1 = 820$ maschi e $N_2 = 750$ femmine. I pesi degli strati sono rispettivamente $w_1 = 0.52$, $w_2 = 0.48$. Inoltre, la statura media dei maschi, nella popolazione, è $\mu_{y1} = 177.00$, mentre quella delle femmine è $\mu_{y2} = 168.26$. La statura media generale è $\mu_y = 172.80$.

Per stimare la statura media μ_y si seleziona un campione ssr di $n = 100$ studenti, di cui si osserva non solo la statura, ma anche il sesso. I dati campionari sono riportati nel *file* `campstature_ssr.txt`. La media campionaria è $\overline{y}_{\mathbf{s}} = 171.98$. Si è in presenza di una sottostima, in quanto $\mu_y = 172.80$.

Ora, nel campione si osservano $n_1 = 47$ maschi e $n_2 = 53$ femmine. Si tratta di un campione un po' "sbilanciato", in quanto contiene il 47% di maschi (contro il 52% nella popolazione) e il 53% di femmine (contro il 48% nella popolazione). Poiché i maschi sono mediamente più alti delle femmine, ci si può attendere che lo sbilanciamento del campione contribuisca a far sottostimare la statura media della popolazione.

Le medie campionarie dei due sottocampioni dei maschi e delle femmine sono rispettivamente pari a

$$\overline{y}_1 = 175.87, \quad \overline{y}_2 = 168.53$$

così che lo stimatore post-stratificato è eguale a

$$\widehat{\mu}_{ps} = 0.52 \times 175.87 + 0.48 \times 168.53 = 172.34.$$

Una parte della sottostima della media campionaria è stata quindi corretta da $\widehat{\mu}_{ps}$. Se si osservano i valori delle medie campionarie e delle numerosità campionarie di strato, è facile accorgersi che la media campionaria $\overline{y}_{\mathbf{s}}$ sottostima μ_y per due ragioni:

1. il sottocampione dei maschi sottostima la corrispondente media di strato, in quanto è $\overline{y}_1 = 175.87$ contro $\mu_{y1} = 177.00$;
2. nel campione vi sono più femmine che maschi (53% contro 47%), mentre nella popolazione accade il contrario (48% contro 52%).

La post-stratificazione elimina in questo caso la causa 2 di sottostima, mentre ovviamente nulla può per la 1. □

Prima di studiare formalmente le proprietà dello stimatore (8.55), sono necessarie alcune precisazioni. Il difetto principale della post-stratificazione riguarda la mancanza di controllo sulla allocazione campionaria $\mathbf{n}$ che può condurre a una o alcune numerosità di strato n_g nulle o molto piccole. In particolare, se qualcuna delle n_g è nulla perde di senso la corrispondente media campionaria $\overline{y}_g$, e quindi lo stimatore post-stratificato non può essere definito come in (8.55). Per ovviare a tale inconveniente si ricorre al *collassamento degli strati* che consiste nel riunificare post-strati campionari vuoti o poveri con post-strati più consistenti. Chiaramente il collassamento implica una *riduzione* del numero di strati operativi, dando vita ad una stratificazione meno fine di quella originaria.

8.6.2 Proprietà elementari dello stimatore post-stratificato

Per studiare le proprietà dello stimatore post-stratificato è in primo luogo necessario studiare le proprietà del numero di unità campionarie dei diversi strati, ossia della variabile aleatoria $\mathbf{n} = (n_1, \ldots, n_M)$. La probabilità di ottenere una data M-pla $(n_1, \ldots, n_M)$ è la somma delle probabilità di tutti i campioni che contengono n_1 unità del primo strato, n_2 unità del secondo strato, …, n_M unità dello strato M-mo. Poiché ciascun campione ha probabilità $1/\binom{N}{n}$, la probabilità della M-pla $\mathbf{n} = (n_1, \ldots, n_M)$ è:

$$\begin{aligned} & Pr(n_1, \ldots, n_M) \\ &= \frac{\#\ \textit{di campioni contenti}\ n_1\ \textit{unità dello strato}\ 1, \ldots, n_M\ \textit{dello strato}\ M}{\binom{N}{n}} \\ &= \frac{\binom{N_1}{n_1} \cdots \binom{N_M}{n_M}}{\binom{N}{n}}. \end{aligned}$$

La probabilità di un campione $\mathbf{s}$ condizionata a $\mathbf{n} = (n_1, \ldots, n_M)$ è quindi

$$\begin{aligned} Pr(\mathbf{s}\,|\mathbf{n}) &= \frac{Pr(\mathbf{s}, \mathbf{n})}{Pr(\mathbf{n})} \\ &= \frac{Pr(\mathbf{s})}{Pr(\mathbf{n})} \\ &= \frac{1}{\binom{N_1}{n_1} \cdots \binom{N_M}{n_M}} \end{aligned} \tag{8.56}$$

se il campione $\mathbf{s}$ contiene n_1 unità dello strato 1, ..., n_M unità dello strato M, mentre è uguale a 0 in caso contrario. Condizionatamente a $\mathbf{n} = (n_1, \ldots, n_M)$, $\mathbf{s}$ è quindi equivalente ad un campione stratificato in cui si selezionano n_1 unità dello strato 1, ..., n_M unità dello strato M. *Attenzione*: il disegno usato per selezionare $\mathbf{s}$ è di tipo ssr. Il condizionamento rispetto a $\mathbf{n}$ implica il "restringersi" ai soli campioni $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$ che contengono n_1 unità dello strato 1, ..., n_M unità dello strato M. Questi *non* sono tutti i possibili campioni che può produrre il disegno ssr. Il condizionamento, in un certo senso, "riduce" lo spazio dei campioni (e ne modifica di conseguenza la probabilità).

Procedendo condizionatamente a $\mathbf{n}$, ed assumendo che le numerosità campionarie di strato siano tutte positive ($n_g > 0$ per ciascuno strato g) è immediato provare la seguente proposizione, dove con i simboli $E_{\mathbf{s}}[\,|\mathbf{n}]$ e $V_{\mathbf{s}}[\,|\mathbf{n}]$ si indicano rispettivamente il valore atteso e la varianza rispetto al disegno campionario, ma *condizionatamente* alla "configurazione" $\mathbf{n} = (n_1, \ldots, n_M)$.

Proposizione 8.6. *Se* $\mathbf{n}$ *è tale che* $n_1 > 0$, ..., $n_M > 0$, *condizionatamente a* $\mathbf{n}$, $\overline{y}_{ps}$ *è uno stimatore corretto della media della popolazione:*

$$E_{\mathbf{s}}\left[\widehat{\mu}_{ps}\,|\mathbf{n}\right] = \mu_y \tag{8.57}$$

e la sua varianza è pari a

$$V_{\mathbf{s}}(\widehat{\mu}_{ps}\,|\mathbf{n}) = \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2. \tag{8.58}$$

Dimostrazione. Per provare la (8.57) basta osservare che, per la (8.56),

$$\begin{aligned} E_{\mathbf{s}}[\widehat{\mu}_{ps}\,|\mathbf{n}] &= E_{\mathbf{s}}\left[\sum_{g=1}^{M} w_g \overline{y}_g \,\middle|\, \mathbf{n}\right] \\ &= \sum_{g=1}^{M} w_g\, E[\overline{y}_g\,|\mathbf{n}] \\ &= \sum_{g=1}^{M} w_g \mu_{yg} \\ &= \mu_y. \end{aligned}$$

Per quanto riguarda la varianza (condizionata) dello stimatore post-stratificato, sempre dalla (8.56) si ricava che

$$\begin{aligned} V_{\mathbf{s}}(\widehat{\mu}_{ps}\,|\mathbf{n}) &= \sum_{g=1}^{M} w_g^2 V_{\mathbf{s}}(\overline{y}_g\,|\mathbf{n}) \\ &= \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 \end{aligned}$$

che è l'usuale formula della varianza per campioni stratificati. □

La stima della varianza *condizionata* (8.58) è molto semplice, in quanto basta sostituire le S^2_{yg} in (8.58) con le corrispondenti varianze campionarie corrette di strato.

Lo studio delle proprietà non condizionate dello stimatore $\widehat{\mu}_{ps}$ richiede alcune complicazioni formali, dovute al modo in cui $\widehat{\mu}_{ps}$ è definito nel caso di campioni che non contengono unità di uno o più strati. La tecnica di collassamento degli strati porta ad uno stimatore *distorto*.

Per capire perché $\widehat{\mu}_{ps}$ è distorto quando non contiene nessuna unità di uno strato, supponiamo che sia $n_1 = 0$, così che nel campione $\mathbf{s}$ non vi sono unità dello strato 1. Supponiamo invece che vi siano unità dello strato 2. Il collassamento dei due strati 1, 2 consiste nel formare un unico "macro-strato" collassato di peso $\widetilde{w} = w_1 + w_2$, e di media $\widetilde{\mu} = \frac{w_1}{w_1+w_2}\,\mu_{y1} + \frac{w_2}{w_1+w_2}\,\mu_{y2}$. Per stimare $\widetilde{\mu}$ viene usata la media campionaria del "macro-strato", che però, essendo $n_1 = 0$, è del tipo

$$\frac{1}{n_2}\sum_{i\in\mathbf{s}_2} y_i$$

e quindi non contiene mai unità dello strato 1. Essa non può quindi essere in alcun modo un stimatore corretto di $\widetilde{\mu}$. Per la distorsione (non condizionata) di $\widehat{\mu}_{ps}$ in un caso speciale si veda l'Esercizio 8.13.

Ad ogni modo, poiché per n "abbastanza grande" la probabilità che una o più delle numerosità campionarie di strato siano nulle diventa sostanzialmente trascurabile, in via approssimata lo stimatore $\widehat{\mu}_{ps}$ è corretto anche non condizionatamente, e sempre in via approssimata si può calcolare la sua varianza. I relativi risultati sono riportati nella successiva Proposizione 8.7.

Proposizione 8.7. *$\widehat{\mu}_{ps}$ è uno stimatore approssimativamente corretto della media della popolazione:*

$$E\,[\widehat{\mu}_{ps}] \approx \mu_y \tag{8.59}$$

e la sua varianza è in via approssimata pari a

$$V(\overline{y}_{ps}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{g=1}^{M} w_g S_g^2 + \frac{1}{n^2}\sum_{g=1}^{M}(1-w_g)S^2_{yg}\,. \tag{8.60}$$

Dimostrazione. Le espressioni approssimate (8.59), (8.60) si ottengono, come detto, considerando trascurabile la probabilità che una o più delle n_g sia pari a 0. Indichiamo con $E_{\mathbf{s}}[\,|\mathbf{n}]$ e $V_{\mathbf{s}}[\,|\mathbf{n}]$ rispettivamente il valore atteso e la varianza rispetto al disegno campionario condizionatamente a $\mathbf{n} = (n_1, \ldots, n_M)$, e con $E_{\mathbf{n}}$, $V_{\mathbf{n}}$ rispettivamente media e varianza rispetto alla distribuzione di probabilità delle numerosità $n_1, \ldots, n_M$. Si ha

$$\begin{aligned}
E[\widehat{\mu}_{ps}] &= E_{\mathbf{n}}\left[E_{\mathbf{s}}\left(\widehat{\mu}_{ps}\,|\,\mathbf{n}\right)\right] \\
&\approx E_{\mathbf{n}}\left[E_{\mathbf{s}}\left(\widehat{\mu}_{ps}\,|\,\mathbf{n}\right)|\,n_1 > 0, \ldots, n_M > 0\right] \\
&= E_{\mathbf{n}}[\mu_y\,|n_1 > 0, \ldots, n_M > 0] \\
&= \mu_y.
\end{aligned}$$

La varianza non condizionata si ricava in modo simile scrivendo

$$\begin{aligned} V(\widehat{\mu}_{ps}) &= E_{\mathbf{n}}\left[V_{\mathbf{s}}(\widehat{\mu}_{ps}\,|\mathbf{n})\right] + V_{\mathbf{n}}\left(E_{\mathbf{s}}[\widehat{\mu}_{ps}\,|\mathbf{n})\right) \\ &\approx E_{\mathbf{n}}\left[V_{\mathbf{s}}(\widehat{\mu}_{ps}\,|\mathbf{n})\,|\,n_1>0,\ldots,n_M>0\right] \\ &\quad + V_{\mathbf{n}}\left(E_{\mathbf{s}}[\widehat{\mu}_{ps}\,|\mathbf{n})\right)|\,n_1>0,\ldots,n_M>0\big)\,. \end{aligned} \tag{8.61}$$

Essendo il secondo termine della (8.61) pari a zero come conseguenza della (8.57), si ottiene

$$\begin{aligned} V(\widehat{\mu}_{ps}) &\approx E_{\mathbf{n}}\left[V_{\mathbf{s}}(\widehat{\mu}_{ps}\,|\mathbf{n})\,|\,n_1>0,\ldots,n_M>0\right] \\ &= E_{\mathbf{n}}\left[\sum_{g=1}^{M} w_g^2\left(\frac{1}{n_g}-\frac{1}{N_g}\right)S_{yg}^2\,\middle|\,n_1>0,\ldots,n_M>0\right] \\ &= \sum_{g=1}^{M} w_g^2\left(E\left[\frac{1}{n_g}\,\middle|\,n_g>0\right]-\frac{1}{N_g}\right)S_{yg}^2 \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\sum_{g=1}^{M} w_g S_g^2 + \frac{1}{n^2}\sum_{g=1}^{M}(1-w_g)S_{yg}^2 \end{aligned}$$

dove per $E(n_g^{-1}\,|n_q>0)$ si utilizza l'approssimazione di Stephan (Cochran, 1977)

$$E\left[\frac{1}{n_g}\,\middle|\,n_g>0\right]\approx\frac{1}{nw_g}-\frac{1}{n^2w_g}+\frac{1}{n^2w_g^2}\,. \qquad \square$$

Nella (8.60) il primo termine coincide con la varianza di un campionamento stratificato con allocazione proporzionale. Il secondo termine è dovuto alla casualità delle numerosità n_g, la quale introduce una ulteriore fonte di variabilità nello stimatore. Tale termine può essere trascurato se la numerosità campionaria n è sufficientemente grande.

L'uso dello post-stratificazione richiede la conoscenza dei pesi $w_g = N_g/N$, desumibili da dati censuari o amministrativi, che generalmente risultano non aggiornati. Se i pesi utilizzati u_g si differenziano dai pesi veri w_g, allora lo stimatore

$$\widehat{\mu}_u = \sum_{g=1}^{M} u_g\,\overline{y}_g \tag{8.62}$$

risulterà diverso dallo stimatore (8.55). Inoltre la distorsione di $\overline{y}_u$ sarà pari a

$$E(\overline{y}_u|\mathbf{n}) - \mu_y = \sum_{g=1}^{M}(u_g - w_g)\mu_{yg}\,. \tag{8.63}$$

Di conseguenza, se i pesi sono poco attendibili nascono dubbi sul reale valore della post-stratificazione poichè in alcune situazioni l'incremento della distorsione potrebbe compensare l'eventuale riduzione nella varianza.

A differenza della stratificazione, nella post-stratificazione le variabili ausiliarie possono essere scelte in relazione alle variabili di interesse i cui parametri si vogliono stimare allo scopo di massimizzare il guadagno in precisione. Ciò si rivela particolarmente utile nelle indagini multiscopo in cui numerose sono le variabili di interesse.

8.6.3 Approfondimenti sugli approcci condizionato e non condizionato

I risultati delle Proposizioni 8.6 e 8.7 evidenziano che il principale problema nello studio della post-stratificazione riguarda il tipo di approccio (condizionato o non condizionato) da utilizzare per valutare le proprietà dello stimatore $\widehat{\mu}_{ps}$. Essendo lo stimatore (8.55) (almeno in via approssimata) corretto in entrambi gli approcci il problema riguarda essenzialmente la scelta della varianza a cui fare riferimento.

Poiché a seguito della post-stratificazione delle unità campionarie il disegno iniziale (ssr) viene apparentemente modificato in un disegno stratificato, la questione teorica da affrontare è se per valutare le proprietà dello stimatore (8.55) bisogna far riferimento al campionamento stratificato o al campionamento semplice senza ripetizione. Nel primo caso si opererà condizionatamente alle numerosità dei post-strati e la varianza di riferimento sarà la (8.58). Nel secondo caso si considererà anche la variabilità dovuta alla casualità di $\mathbf{n}$ e la varianza di riferimento sarà la (8.60).

A favore dell'approccio condizionato c'è l'osservazione che una volta estratto il campione, la distribuzione dei campioni di numerosità n con allocazione $\mathbf{n}$ diversa da quella osservata è irrilevante per la stima, e di conseguenza la variabilità campionaria di $\mathbf{n}$ non deve essere considerata perché l'elemento condizionante è solo strumento e non oggetto diretto di inferenza.

Inoltre poiché la varianza non condizionata è costante qualunque sia l'allocazione campionaria, dall'utilizzo della (8.60) si ricaverebbe la stessa precisione indipendentemente dalla configurazione di $\mathbf{n}$. Ciò è chiaramente controintuitivo.

A favore dell'approccio non condizionato vi è invece l'osservazione che la varianza non condizionata (8.60) riveste un ruolo decisivo quando, nella progettazione di nuove indagini, è necessario effettuare una scelta tra strategie campionarie alternative.

A seguito di questa affermazione, procediamo a confrontare, a parità di numerosità campionaria n, la *performance* di $\widehat{\mu}_{ps}$ con quella della media campionaria usata con il disegno ssr. Prima della estrazione del campione il confronto deve essere effettuato utilizzando l'approccio non condizionato. Dalla formula approssimata (8.60) risulta chiaramente che per campioni grandi ci si può attendere dalla post-stratificazione un guadagno di efficienza rispetto al campionamento casuale semplice. Tale guadagno di efficienza è comparabile con quello derivante da una stratificazione con allocazione proporzionale.

Dopo l'estrazione del campione il confronto deve essere effettuato condizionatamente all'allocazione campionaria realizzatasi. A tale scopo occorre innanzitutto valutare le proprietà dello stimatore media campionaria all'interno dell'approccio post-stratificato. Si osservi che la media campionaria può essere scritta nel seguente modo

$$\overline{y}_{\mathbf{s}} = \frac{1}{n} \sum_{g=1}^{M} \sum_{i \in \mathbf{s}_g} y_{gj} = \sum_{g=1}^{M} \left(\frac{n_g}{n} \right) \overline{y}_g.$$

Condizionatamente a $\mathbf{n}$, si ha

$$E\left[\overline{y}_{\mathbf{s}} \,|\mathbf{n} \right] = \sum_{g=1}^{M} \frac{n_g}{n} \mu_{yg}$$

per cui lo stimatore $\overline{y}_{\mathbf{s}}$, che è non distorto sotto un campionamento casuale ssr, diventa distorto all'interno dell'approccio post-stratificato. La sua distorsione è pari a

$$E\left[\overline{y}_{\mathbf{s}} \,|\mathbf{n} \right] - \mu_y = - \sum_{g=1}^{M} \mu_{yg} \left(\frac{N_g}{N} - \frac{n_g}{n} \right).$$

La distorsione della media campionaria è nulla se vale una delle due seguenti affermazioni:

- il campione è distribuito proporzionalmente tra gli strati: $n_g/n = N_g/N$ per ogni $g = 1, \ldots, M$ (sotto queste condizioni l'errore quadratico medio si riduce alla varianza condizionata (8.58));
- μ_{yg} è costante in tutti i post-strati campionari.

Poiché la media campionaria $\overline{y}_{\mathbf{s}}$ risulta distorta, bisogna procedere al calcolo del suo errore quadratico medio condizionato:

$$\begin{aligned} MSE(\overline{y}_{\mathbf{s}} \,|\mathbf{n}) &= V(\overline{y}_{\mathbf{s}} \,|\mathbf{n}) + (E(\overline{\mathbf{s}}_s \,|\mathbf{n}) - \mu_y)^2 \\ &= \sum_{g=1}^{M} \left(\frac{n_g}{n} \right)^2 \left(1 - \frac{n_g}{N_g} \right) \frac{S_{yg}^2}{n_g} + \left\{ \sum_{g=1}^{M} \mu_{yg} \left(\frac{N_g}{N} - \frac{n_g}{n} \right) \right\}^2 . \quad (8.64) \end{aligned}$$

Confrontando l'errore quadratico medio $MSE(\overline{y}_{\mathbf{s}} \,|\mathbf{n})$ con la varianza dello stimatore post-stratificato (8.58), si ha che

$$\begin{aligned} MSE(\overline{y}_{\mathbf{s}} \,|\mathbf{n}) - V(\widehat{\mu}_{ps} \,|\mathbf{n}) &= \left\{ \sum_{g=1}^{M} \mu_{yg} \left(\frac{N_g}{N} - \frac{n_g}{n} \right) \right\}^2 \\ &\quad + \sum_{g=1}^{M} \left\{ \left(\frac{n_g}{n} \right)^2 - \left(\frac{N_g}{N} \right)^2 \right\} \left(1 - \frac{n_g}{N_g} \right) \frac{S_{yg}^2}{n_g}. \quad (8.65) \end{aligned}$$

Il segno della (8.65) dipende dalla allocazione campionaria $\mathbf{n}$, dalle medie e dalle varianze dei post-strati. In conclusione non possiamo affermare che uno stimatore è uniformemente migliore dell'altro; ogni situazione deve essere esaminata separatamente. Questa conclusione è valida per tutti i campioni stratificati, siano essi pre-stratificati o post-stratificati. In molti casi la post-stratificazione fornisce risultati migliori del disegno ssr a parità di numerosità campionaria. Inoltre, va sottolineato che la post-stratificazione protegge le inferenze dall'utilizzo di campioni "sbilanciati" caratterizzati da configurazioni campionarie $\mathbf{n}$ estreme. In particolare, il ricorso alla post-stratificazione può ridurre la distorsione dovuta sia alla non rappresentatività del campione sia alla presenza di mancate risposte.

Esercizi

8.1. Verificare che la soluzione del problema di minimo (8.17) è la seguente: $D_g = K/M$ per ogni $g = 1, \ldots, M$.

Suggerimento. Usare la tecnica dei moltiplicatori di Lagrange, con funzione Lagrangiana del tipo $\mathcal{L}(D_1, \ldots, D_M; \lambda) = \sum_g D_g^2 - 2\,\lambda\,(\sum_g D_g - K)$.

8.2. (*Stratificazione ottimale con allocazione proporzionale)* Nella Sezione 8.1.1 si è supposto che l'allocazione delle unità campionarie negli strati sia quella di Neyman. Supponendo di usare un'allocazione proporzionale, provare che i valori di $l_1, \ldots, l_{M-1}$ che minimizzano $V(\widehat{\mu}_{str};\ prop)$ soddisfano le relazioni:

$$l_g = \frac{\mu_{yg} + \mu_{y\,g+1}}{2},\ \ g = 1, \ldots, M-1.$$

Suggerimento. Bisogna determinare $l_1, \ldots, l_{M-1}$ in modo da rendere minima la $\sum_g w_g S_{yg}^2$. Usando la stessa notazione della Sezione 8.1.1, si ha

$$\frac{\partial\, w_g\, S_{yg}^2}{\partial\, l_g} = l_g^2\, f_Y(l_g) - 2\, l_g\, \mu_{yg}\, f_Y(l_g) + f_Y(l_g)\, \mu_{yg}^2$$

$$\frac{\partial\, w_{g+1}\, S_{y\,g+1}^2}{\partial\, l_g} = -l_g^2\, f_Y(l_g) + 2\, l_g\, \mu_{y\,g+1}\, f_Y(l_g) - f_Y(l_g)\, \mu_{y\,g+1}^2$$

da cui

$$\frac{\partial}{\partial\, l_g}\left(\sum_g w_g\, S_{yg}^2\right) = f_Y(l_g)\,(\mu_{yg} - \mu_{y\,g+1})\,\{2\, l_g + (\mu_{yg} + \mu_{y\,g+1})\},$$

$g = 1, \ldots, M$.

8.3. Usando le ipotesi e la notazione della Sezione 8.1.3, provare che i valori di $t_1, \ldots, t_{M-1}$ che minimizzano la $\sum_g w_g\, S_{yg}^2$ sono ottenuti, in via approssimata, in modo tale che le quantità

$$\int_{t_{g-1}}^{t_g} \sqrt[3]{f_X(x)}\, dx$$

siano costanti.

Suggerimento. Se $f_X(x) \approx f_{xg}$ nello strato g-mo, si ha

$$\sum_g w_g\, S_{xg}^2 \approx \frac{1}{12} \sum_g \left(\int_{t_{g-1}}^{t_g} \sqrt[3]{f_X(x)}\, dx \right)^3 .$$

8.4. Con riferimento alla popolazione di 385 aziende dell'Esempio 8.1, suddividere tale popolazione in $M = 2$ strati in modo che la $\sum_g w_g\, S_{xg}$ sia minima. Confrontare il risultato con quello che si ottiene usando (a) la regola cum $\sqrt{f}$, e (b) la regola di Ekman.

8.5. Verificare che, nelle ipotesi della Sezione 8.2, se il carattere $\mathcal{X}$ ha distribuzione uniforme in $(x_{min},\, x_{max})$ e se gli strati hanno tutti la stessa ampiezza $(x_{max} - x_{min})/M$, allora vale la relazione $S_{xg}^2 = S_x^2/M^2$.

8.6. Verificare che vale la relazione (8.28).

8.7. Verificare che la soluzione del problema di ottimo vincolato (8.30) è data dalla (8.31).

Suggerimento. Usare la tecnica dei moltiplicatori di Lagrange, con funzione Lagrangiana del tipo: $\mathcal{L}(n_1,\, \ldots,\, n_M\ \lambda) = \sum_{g=1}^M \frac{w_g^2 V_g^2}{n_g} + \lambda \left(\sum_{g=1}^M n_g - n \right)$.

8.8. Provare che vale la relazione (8.38).

Suggerimento. $B(\widehat{\mu}_{q\,com}) = -(E[\widehat{R}_{str}\, \widehat{\mu}_{q\,com}] - E[\widehat{R}_{str}]\, E[\widehat{\mu}_{q\,com}])$.

8.9. Provare che (8.45) è uno stimatore corretto di μ_y, e che la sua varianza è pari alla (8.46).

Suggerimento. $V(\widehat{\mu}_{gd\,sep} = \sum_{g=1}^M w_g^2\, V(\overline{y}_g - c_g\, \overline{x}_g) = \sum_{g=1}^M w_g^2\, \{V(\overline{y}_g) + c_g^2\, V(\overline{x}_g) - 2\, c_g\, C(\overline{y}_g,\, \overline{x}_g)\}$.

8.10. Verificare che i valori di $c_1,\, \ldots,\, c_M$ che minimizzano la (8.46) sono del tipo $c_g = b_{y/x\,g}$, $g = 1,\, \ldots,\, M$.

Suggerimento. Il valore di c_g che rende minima la $S_{yg}^2 + c_g^2\, S_{xg}^2 - 2\, c_g\, S_{xy\,g}$ è pari a $S_{xy\,g}/S_{xg}^2 = b_{y/x\,g}$.

8.11. Provare che (8.48) è uno stimatore corretto di μ_y, e che la sua varianza è pari alla (8.49).

Suggerimento. $V(\widehat{\mu}_{gd\,c} = \sum_{g=1}^M w_g^2\, V(\overline{y}_g - c\,\overline{x}_g) = \sum_{g=1}^M w_g^2\, \{V(\overline{y}_g) + c^2\, V(\overline{x}_g) - 2\, c\, C(\overline{y}_g,\, \overline{x}_g)\}$.

8.12. Verificare che il valore di c che rende minima la (8.49) è (8.50).

Suggerimento. Derivare la (8.49) rispetto a c e annullare la derivata.

8.13. Si consideri una popolazione suddivisa in $M = 2$ strati, e sia $\mathbf{s}$ un campione ssr di n unità. Con la solita notazione, si consideri lo stimatore

$$\widehat{\mu}_{ps} = \begin{cases} w_1 \overline{y}_1 + w_2 \overline{y}_2 \; se\; n_1 > 0,\; n_2 > 0 \\ \overline{y}_1 \; se\; n_2 = 0 \\ \overline{y}_2 \; se\; n_1 = 0 \end{cases} .$$

a. Provare che

$$E[\widehat{\mu}_{ps}] = \mu_{y1}\, Pr(n_2 = 0) + \mu_{y2}\, Pr(n_1 = 0) + \mu_y\, (1 - Ps(n_1 = 0) - Pr(n_2 = 0)).$$

b. Verificare che

$$Pr(n_1 = 0) = \frac{\binom{N_2}{n}}{\binom{N}{n}}, \quad Pr(n_2 = 0) = \frac{\binom{N_1}{n}}{\binom{N}{n}}.$$

c. Concludere dai punti precedenti che $E[\widehat{\mu}_{ps}] \neq \mu_y$.

9

Disegno campionario a grappolo con uguali probabilità di selezione

9.1 La nozione di "grappolo": aspetti di base e notazione

Nei disegni campionari finora presi in esame si è sempre assunto che è data una lista di unità elementari (unità di rilevazione) della popolazione, e che la procedura di selezione del campione agisca direttamente su tali unità. Nella terminologia introdotta nella Sezione 1.5, le unità di campionamento coincidono con le unità di osservazione.

In molte rilevazioni campionarie questo non accade. Molto frequente è invece il caso in cui le unità di campionamento sono aggregati, *grappoli* di unità di rilevazione. Nel seguito useremo come equivalenti i termini "grappoli" e "unità primarie" .

Un caso molto importante in cui la considerazione di grappoli sorge in maniera "naturale" è quello in cui non si ha una lista delle unità elementari (di rilevazione) della popolazione, ma solo una lista di unità primarie. Si supponga ad esempio che una casa produttrice di giocattoli voglia effettuare una indagine campionaria per verificare il livello di gradimento di alcuni suoi prodotti. La popolazione obiettivo è quella dei bambini della fascia di età per la quale l'azienda produce i propri giocattoli. Ora, in generale non esiste, o comunque non è accessibile, una lista di bambini da cui selezionare un campione. Vi sono però altre possibilità. Ad esempio, si potrebbe pensare di selezionare un campione di famiglie, e poi di intervistare i bambini delle famiglie selezionate. In alternativa, si potrebbe anche selezionare un campione di classi scolastiche, e intervistare tutti i bambini delle classi selezionate. Nel primo caso le unità di campionamento sono le famiglie, ciascuna delle quali include un certo numero di bambini (eventualmente nessuno), e che quindi può essere vista come un grappolo di bambini. Nel secondo caso le unità di campionamento sono classi scolastiche, ciascuna delle quali, similmente, può essere considerata come un grappolo di unità elementari. Questo semplice esempio si presta a diverse considerazioni.

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_9,

1. Il considerare grappoli consente di ovviare all'inconveniente dovuto alla non disponibilità di una lista di unità elementari. Tutto ciò di cui si ha bisogno è una lista di grappoli, spesso non difficile da reperire.
2. Per una stessa popolazione possono esistere diversi tipi di grappoli di unità elementari. Il tipo di grappolo usato come unità di campionamento dipende da elementi quali la disponibilità di una lista, la facilità di accesso e il relativo costo, etc.

L'idea di base del disegno campionario a grappolo con uguali probabilità di selezione, in estrema sintesi, è molto semplice: si seleziona, con disegno ssr, un *campione di grappoli* e si osservano tutte le unità elementari dei grappoli campionati.

9.1.1 Simbologia utilizzata

D'ora in avanti supporremo che nella popolazione vi siano M grappoli, formati rispettivamente da N_1, N_2, ..., N_M unità elementari. Esattamente come nel caso di popolazioni ripartite in strati, ogni unità elementare è individuata da una doppia etichetta (g, i), con:

- $g(= 1, \ldots, M)$ è il grappolo a cui l'unità appartiene;
- $i(= 1, \ldots, N_g)$ indica l'unità nell'ambito del grappolo di appartenenza.

Indicheremo poi con $w_g = N_g/N$ il peso del grappolo g-mo ($g = 1, \ldots, M$).

La notazione introdotta è identica a quella usata per gli strati. Dal punto di vista *formale*, in effetti, non vi è praticamente nessuna differenza tra strati e grappoli. Sia gli strati che i grappoli sono insiemi, aggregati di unità elementari, che costituiscono una partizione della popolazione (ogni unità elementare appartiene ad uno e un solo grappolo/strato). Dal punto di vista sostanziale, statistico, le differenze sono invece enormi. Nel disegno campionario stratificato le unità di campionamento sono quelle elementari, che vengono selezionate separatamente per ciascuno strato. È quindi necessario disporre, per ogni strato, di una lista di unità elementari. Nel caso del disegno a grappolo, invece, le unità di campionamento sono i grappoli. Non è quindi necessario disporre, *a priori*, di una lista delle unità elementari nei diversi grappoli. Ciò di cui si ha bisogno è: (a) una lista dei grappoli da cui la popolazione è formata; (b) una lista delle unità elementari dei soli grappoli campionati.

Sia y_{gi} la modalità dell'unità i $(= 1, \ldots, N_g)$ del grappolo g $(= 1, \ldots, M)$, e con

$$\mu_{yg} = \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi}, \quad S^2_{yg} = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2 ; \; g = 1, \ldots, M$$

rispettivamente la media e la varianza corretta del carattere di interesse $\mathcal{Y}$ nel grappolo g-mo. Per semplificare la notazione, per ogni grappolo g poniamo

poi

$$z_g = M\, w_g\, \mu_{yg} = \frac{M}{N} \sum_{i=1}^{N_g} y_{gi}\,;\ g = 1, \ldots, M. \tag{9.1}$$

Per quanto riguarda la media della popolazione, vale la seguente relazione:

$$\begin{aligned} \mu_y &= \sum_{g=1}^{M} w_g\, \mu_{yg} \\ &= \frac{1}{M} \sum_{g=1}^{M} M\, w_g\, \mu_{yg} \\ &= \frac{1}{M} \sum_{g=1}^{M} z_g. \end{aligned} \tag{9.2}$$

La (9.2) mette in evidenza un fatto molto importante: la media μ_y del carattere di interesse $\mathcal{Y}$ nella popolazione può essere espressa come una semplice media delle quantità $z_1, \ldots, z_M$ (9.1).

9.1.2 Il disegno campionario a grappolo

Come già anticipato, l'idea di base del disegno campionario a grappolo è molto semplice: si seleziona, mediante campionamento ssr, un campione $\mathbf{g}_m$ di m degli M grappoli totali, e si osservano le modalità di tutte le unità elementari dei grappoli campionati. Formalmente, lo spazio dei campioni è l'insieme $\mathcal{C}_{M,m}$ di tutte le combinazioni senza ripetizione di m degli M grappoli. Ciascuna di tali combinazioni ha probabilità

$$p(\mathbf{g}_m) = \frac{1}{\binom{M}{m}} \quad \textit{per ogni} \ \ \mathbf{g}_m \in \mathcal{C}_{M,m}.$$

Sempre sul piano formale, questo significa che i nostri dati campionari sono le modalità

$$y_{gi}, \quad \textit{per ciascuna unità } i = 1, \ldots, N_g \textit{ e per tutti i grappoli } g \textit{ in } \mathbf{g}_m.$$

Sulla base dei dati campionari è possibile calcolare, in particolare, le medie, le varianze e le quantità z_g (9.1) per tutti i grappoli del campione $\mathbf{g}_m$.

Il disegno a grappolo è per molti aspetti simile al disegno ssr. La principale differenza consiste nel fatto che nel disegno ssr le unità di campionamento coincidono con quelle elementari, mentre nel disegno a grappolo si campionano grappoli di unità elementari. Lo schema qui di seguito riportato mette in luce tale corrispondenza.

	Disegno ssr	*Disegno a grappolo*
Unità di campionamento	Unità elementari i	Grappoli g
Numero totale di unità di campionamento	N	M
Quantità da stimare	$\frac{1}{N}\sum_{i=1}^{N} y_i$	$\frac{1}{M}\sum_{g=1}^{M} z_g$
Numero di unità campionate	n	m
Spazio dei campioni	$\mathcal{C}_{N,n}$	$\mathcal{C}_{M,m}$
Probabilità dei campioni	$1/\binom{N}{n}$	$1/\binom{M}{m}$
Quantità osservate nel campione	y_i	z_g

9.2 Stima della media della popolazione

Sulla base della (9.2) e delle corrispondenze con il disegno ssr evidenziate nella sezione precedente, l'intuizione suggerisce di stimare la media μ_y della popolazione con la media campionaria delle z_g. Si ha in questo modo lo *stimatore a grappolo*:

$$\begin{aligned}\widehat{\mu}_{gr} &= \textit{media campionaria delle } z_g \\ &= \frac{1}{m}\sum_{g\in\mathbf{g}_m} z_g \\ &= \frac{1}{m}\sum_{g\in\mathbf{g}_m} M\, w_g\, \mu_{yg}. \end{aligned} \tag{9.3}$$

Le proprietà dello stimatore (9.3) sono studiate nella Proposizione 9.1. Per comodità di notazione, indichiamo con

$$S_b^2 = \frac{1}{M-1}\sum_{g=1}^{M}(z_g-\mu_y)^2 \tag{9.4}$$

la varianza delle quantità z_g nella popolazione (corretta con un denominatore $M-1$ anziché M).

Proposizione 9.1. *Se il disegno campionario è a grappolo, $\widehat{\mu}_{gr}$ è uno stimatore corretto della media della popolazione:*

$$E\left[\widehat{\mu}_{gr}\right] = \mu_y \tag{9.5}$$

e la sua varianza è pari a

$$V\left(\widehat{\mu}_{gr}\right) = \left(\frac{1}{m}-\frac{1}{M}\right) S_b^2. \tag{9.6}$$

Dimostrazione. È sufficiente tenere conto che $\widehat{\mu}_{gr}$ è la media campionaria delle z_g, e che il campione $\mathbf{g}_m$ di grappoli è selezionato con disegno semplice senza ripetizione. □

La quantità S_b^2 è la varianza (corretta) delle quantità $z_1, \ldots, z_M$ nella popolazione dei grappoli. Detto

$$T_g = \sum_{i=1}^{N_g} y_{gi} = N_g \, \mu_{yg}, \;\; g = 1, \ldots, M$$

il totale (l'ammontare) del carattere $\mathcal{Y}$ nel grappolo g-mo, valgono le relazioni

$$z_g = \frac{M}{N} N_g \, \mu_{yg} = \frac{M}{N} T_g, \;\; g = 1, \ldots, M$$

dalle quali discende che

$$S_b^2 = Varianza \; di \; z_1, \ldots, z_m \; = \; \left(\frac{M}{N}\right)^2 \times (Varianza \; di \; T_1, \ldots, T_m) \, . \quad (9.7)$$

Le due relazioni (9.6) e (9.7) ci dicono che l'errore quadratico medio (e quindi l'efficienza) dello stimatore $\widehat{\mu}_{gr}$ dipende essenzialmente da due elementi:

1. la variabilità dei totali dei grappoli;
2. il numero di grappoli campionati.

La 1 è tutto sommato ovvia: quanti più grappoli si campionano, tanto migliore è la stima della media della popolazione che si ottiene. Molto più interessante è invece la 2. Il termine S_b^2 che determina la varianza di $\widehat{\mu}_{gr}$ è proporzionale alla varianza dei totali $T_1, \ldots, T_M$. Quanto più $T_1, \ldots, T_M$ sono "simili" tra loro, tanto più piccola è S_b^2, e quindi tanto più efficiente è lo stimatore $\widehat{\mu}_{gr}$. In altri termini, *la coppia* (*disegno campionario a grappolo, stimatore* $\widehat{\mu}_{gr}$) *fornisce stime tanto migliori quanto più bassa è la variabilità dei totali dei grappoli.*

Per quanto riguarda la stima della varianza (9.6), si possono ancora usare risultati ben noti per il disegno ssr. Le stesse idee su cui si basa la stima della varianza della media campionaria nel disegno ssr portano infatti a considerare la varianza campionaria corretta delle z_g:

$$\widehat{s}_b^2 = \frac{1}{m-1} \sum_{g \in \mathbf{g}_m} (z_g - \widehat{\mu}_{gr})^2. \quad (9.8)$$

Come semplice adattamento della Proposizione 3.3, è immediato verificare che lo stimatore (9.8) è uno stimatore corretto della (9.4).

Proposizione 9.2. *Se il disegno campionario è a grappolo,* $\widehat{s}_b^2$ *è uno stimatore corretto della varianza* S_b^2*:*

$$E\left[\widehat{s}_b^2\right] = S_b^2.$$

Infine, dalla Proposizione 9.2 è pressoché immediato trarre uno stimatore corretto di $V(\widehat{\mu}_{gr})$.

Proposizione 9.3. *Se il disegno campionario è a grappolo, la quantità*

$$\widehat{V}_{gr} = \left(\frac{1}{m} - \frac{1}{M}\right) \widehat{s}_b^2 \tag{9.9}$$

è uno stimatore corretto della varianza $V(\widehat{\mu}_{gr})$ (9.6).

Esempio 9.1. Il comune di Statlandia non possiede anagrafe, per cui non esiste una lista delle famiglie (e tantomeno degli individui) in esso residenti. Tutto ciò che è noto è che a Statlandia vi sono in totale 128 palazzi, ciascuno di 8 appartamenti. I palazzi sono identificati da un numero intero compreso tra 1 e 128; gli appartamenti di ciascun palazzo da un numero intero compreso tra 1 e 8. Poiché ciascuna famiglia vive in uno e un solo appartamento, a Statlandia risiedono in totale 1024 famiglie. Ciò fornisce, in linea di principio, anche un modo per etichettare le famiglie. Infatti, ciascuna famiglia è identificata da una doppia etichetta, del tipo (numero del palazzo, numero di appartamento nel palazzo).

Nel *file* `fam2051.txt` sono riportati i dati relativi alle famiglie del comune di Statlandia, e rilevati il 30 giugno 2051. In totale, sono riportati i valori di 27 variabili. Il significato di ciascuna variabile, e la relativa codifica, è riportato nel *file* `istruzioni_fam2051.txt`.

Supponiamo di essere interessati alla stima del reddito medio da lavoro (nell'anno 2050) delle famiglie di Statlandia, le quali sono quindi le unità elementari. Poiché non si dispone di una lista delle famiglie, effettuare un campionamento ssr direttamente su di esse è impossibile. D'altra parte, per la nostra indagine si può utilizzare la lista dei 128 palazzi, i quali possono essere visti come *grappoli* di famiglie. Ciascun grappolo, nel presente caso, è composto da 8 famiglie.

Un'idea molto naturale è quella di effettuare un campionamento a grappolo, selezionando un campione ssr di grappoli (palazzi), e osservando tutte le famiglie che risiedono in ciascuno dei grappoli campionati. Nel caso in esame vi sono $M = 128$ grappoli, ciascuno composto da $L = 8$ unità elementari di osservazione (famiglie). I redditi totali relativi ad un campione di $m = 9$ grappoli sono riportati nel *file* `camp91.txt`. In Tabella 9.1 sono riportate le grandezze necessarie per costruire lo stimatore $\widehat{\mu}_{gr}$ e per stimare la sua varianza.

Come stima del reddito medio da lavoro della popolazione si ha quindi la seguente:

$$\widehat{\mu}_{gr} = \frac{1}{9}\left(z_3 + z_{11} + \cdots + z_{126}\right) = 54537.5.$$

Per quanto riguarda invece la stima della varianza di $\widehat{\mu}_{gr}$, essa assume il valore:

$$\widehat{V}_{gr} = \left(\frac{1}{9} - \frac{1}{128}\right) \frac{1}{8} \sum (z_g - \widehat{\mu}_{gr})^2 = 211288510.9\,. \qquad \square$$

Tabella 9.1 Costruzione degli stimatori $\widehat{\mu}_{gr}$ e $\widehat{V}_{gr}$

Grappolo	z_g	$(z_g - \widehat{\mu}_{gr})^2$
3	128415.6	5457977354
11	127251.1	5287271261
32	61240.0	44933560.6
52	50836.1	13700176.9
64	43987.4	111305137.5
79	33500.0	442576406.3
87	25725.0	830160156.3
94	19881.3	1201055664
126	0.0	2974338906

La costruzione di intervalli di confidenza, infine, si basa su argomenti del tutto simili a quelli usati nei capitoli precedenti. Se il numero m di grappoli campionati è sufficientemente grande, lo stimatore $\widehat{\mu}_{gr}$ ha distribuzione approssimata di tipo normale, con media μ_y e varianza $V(\widehat{\mu}_{gr})$. Ragionando esattamente come nei capitoli precedenti, e sostituendo l'incognita $V(\widehat{\mu}_{gr})$ con la sua stima (9.9), si ha che la distribuzione di probabilità di

$$\frac{\widehat{\mu}_{gr} - \mu_y}{\sqrt{\widehat{V}_{gr}}} \tag{9.10}$$

ha distribuzione approssimata di tipo normale standard. Detto pertanto, come al solito, z_α il quantile di ordine α della distribuzione normale standard, è immediato verificare che

$$\left[\widehat{\mu}_{gr} - z_{\alpha/2}\sqrt{\widehat{V}_{gr}},\ \widehat{\mu}_{gr} + z_{\alpha/2}\sqrt{\widehat{V}_{gr}}\right] \tag{9.11}$$

è un intervallo di confidenza per μ_y, al livello approssimato $1 - \alpha$.

Esempio 9.2. Consideriamo ancora l'Esempio 9.1. Essendo $m = 9$, il numero di grappoli campionati è molto probabilmente troppo piccolo perché la (9.10) abbia, con buona approssimazione, distribuzione normale standard. Tuttavia, a puro titolo di esempio numerico costruiamo l'intervallo (9.11) al livello di confidenza 0.95. Essendo $\sqrt{\widehat{V}(\widehat{\mu}_{gr})} = 14535.8$ e $z_{0.025} = 1.96$, si ha che l'intervallo

$$[54537.5 - 1.96\,14535.8,\ 54537.5 + 1.96\,14535.8] = [26049.5,\ 83025.5]$$

è un intervallo di confidenza approssimato per μ_y al livello 0.95. Come si vede, si tratta di un intervallo estremamente ampio, principalmente a causa del basso numero di grappoli campionati. Questo, tra l'altro, rende difficilmente difendibile l'approssimazione normale usata per la (9.10). □

9.3 Un importante caso speciale: grappoli della stessa dimensione

Un caso speciale molto importante è quello in cui gli M grappoli sono tutti formati dallo stesso numero L di unità:

$$N_1 = N_2 = \cdots = N_M = L.$$

In questo caso si ha infatti $N = M\,L$, per cui i pesi w_g sono tutti uguali

$$w_g = \frac{L}{N} = \frac{1}{M}, \quad g = 1, \ldots, M.$$

Di conseguenza, le quantità z_g (9.1) si riducono alle medie dei grappoli:

$$z_g = \mu_{yg}, \quad g = 1, \ldots, M$$

e lo stimatore $\widehat{\mu}_{gr}$ diviene la media campionaria delle medie dei grappoli:

$$\widehat{\mu}_{gr} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} \mu_{yg}$$

la quale coincide ovviamente con la media campionaria

$$\frac{1}{mL} \sum_{g \in \mathbf{g}_m} \sum_{i=1}^{N_g} y_{gi}.$$

La varianza di $\widehat{\mu}_{gr}$ si presta a considerazioni di interesse. In primo luogo, dalla Proposizione 1.1 e dalla (1.4), tenendo conto che tutte le quantità w_g sono uguali a $1/M$, si ha la relazione:

$$\begin{aligned} \sum_g \sum_i (y_{gi} - \mu_y)^2 &= \sum_g \sum_i (\mu_{yg} - \mu_y)^2 + \sum_g \sum_i (y_{gi} - \mu_{yg})^2 \\ &= L \sum_g (\mu_{yg} - \mu_y)^2 + \sum_g \sum_i (y_{gi} - \mu_{yg})^2 . \end{aligned} \tag{9.12}$$

La quantità

$$D_y^2 = \sum_g \sum_i (y_{gi} - \mu_y)^2$$

è la *devianza totale* per la popolazione. Invece, le due quantità

$$D_b^2 = L \sum_g (\mu_{yg} - \mu_y)^2, \quad D_w^2 = \sum_g \sum_i (y_{gi} - \mu_{yg})^2$$

sono rispettivamente la *devianza tra i grappoli* e la *devianza nei grappoli*. La (9.12) si può riscrivere come

$$D_y^2 = D_b^2 + D_w^2.$$

Dalla relazione

$$S_b^2 = \frac{1}{L\,(M-1)}\,D_b^2$$

e dalla (9.6) si deduce facilmente che l'efficienza dello stimatore $\widehat{\mu}_{gr}$ dipende dalla devianza tra i grappoli: quanto più piccola è D_b^2, tanto più piccola è $V(\widehat{\mu}_{gr})$. Si ha quindi l'esatto contrario di quanto accade per il campionamento stratificato, in cui l'efficienza dello stimatore $\widehat{\mu}_{str}$ è tanto più elevata quanto più piccola è la devianza negli strati (e quindi quanto più grande è la varianza tra gli strati).

In altre parole nel campionamento a grappolo la situazione ideale è che tutti i grappoli in cui risulta suddivisa la popolazione siano più eterogenei possibile al loro interno. Al limite, se ciascun grappolo fosse una copia ridotta della popolazione allora sarebbe sufficiente estrarne uno solo per avere la stessa informazione che si otterrebbe da una indagine completa.

Definiamo il *coefficiente di correlazione intra-classi* come:

$$\rho_{ic} = 1 - \frac{L}{L-1}\,\frac{D_w^2}{D_y^2}\,. \tag{9.13}$$

Poiché, per la (9.13), è $0 \leqslant \frac{D_w^2}{D_y^2} \leqslant 1$, si avrà

$$-\frac{1}{L-1} \leqslant \rho_{ic} \leqslant 1\,. \tag{9.14}$$

In particolare, ρ_{ic} assume il suo massimo valore, 1, quando $D_w^2 = 0$, ossia quando $D_b^2 = D_y^2$. L'uguaglianza $D_w^2 = 0$ significa che non vi è variabilità nei grappoli, ossia che tutte le unità elementari di uno stesso grappolo hanno lo stesso valore della variabile $\mathcal{Y}$ di interesse. È questo il caso di *massima omogeneità nei* grappoli. All'opposto, ρ_{ic} assume il suo valore minimo, $-1/(L-1)$, quando $D_w^2 = D_y^2$, ossia quando $D_b^2 = 0$. Quest'ultima uguaglianza ha luogo quando le medie dei grappoli sono tutte uguali, il che corrisponde alla *minima omogeneità* nei grappoli stessi.

In forza delle considerazioni sopra riportate, il coefficiente di correlazione intra-classi può essere considerato come una misura dell'omogeneità dei grappoli da cui è formata la popolazione. Un'espressione alternativa per ρ_{ic} è riportata nell'Esercizio 9.3.

Per quanto riguarda il termine S_b^2 (9.4), è facile provare (Esercizio 9.2) che se i grappoli hanno tutti la stessa numerosità L vale la relazione

$$S_b^2 = \frac{M\,L-1}{(M-1)\,L^2}\,S_y^2\,(1+(L-1)\,\rho_{ic}) \tag{9.15}$$

in cui

$$S_y^2 = \frac{1}{M\,L-1}\sum_g\,\sum_i (y_{gi}-\mu_y)^2 = \frac{1}{M\,L-1}\,D_y^2$$

è la varianza corretta della popolazione. Sulla base della (9.15) è facile provare il seguente risultato.

Proposizione 9.4. *Se il disegno campionario è a grappolo, e se i grappoli sono tutti costituiti dallo stesso numero L di unità elementari, si ha*

$$V(\widehat{\mu}_{gr}) = \left(\frac{1}{m\,L} - \frac{1}{M\,L}\right) \frac{M\,L-1}{(M-1)\,L}\, S_y^2\,(1+(L-1)\,\rho_{ic}). \tag{9.16}$$

Dimostrazione. È sufficiente tener conto che, per le (9.15), (9.6) si ha:

$$\begin{aligned} V(\widehat{\mu}_{gr}) &= \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 \\ &= \left(\frac{1}{m} - \frac{1}{M}\right) \frac{M\,L-1}{(M-1)\,L^2}\, S_y^2\,(1+(L-1)\,\rho_{ic}) \\ &= \left(\frac{1}{m\,L} - \frac{1}{M\,L}\right) \frac{M\,L-1}{(M-1)\,L}\, S_y^2\,(1+(L-1)\,\rho_{ic}). \end{aligned}$$

□

Nel caso in cui L sia piccolo rispetto a $N = M\,L$, come spesso accade in pratica, si ha in via approssimata $M\,L-1 \approx (M-1)\,L$, e quindi dalla (9.16) si trae la relazione approssimata

$$V(\widehat{\mu}_{gr}) \approx \left(\frac{1}{m\,L} - \frac{1}{M\,L}\right) S_y^2\,(1+(L-1)\,\rho_{ic}). \tag{9.17}$$

L'interesse dell'espressione (9.17) sta nel fatto che essa permette di confrontare, in termini molto semplici, l'efficienza della coppia (*disegno ssr*, *media campionaria*) con quella della coppia (*disegno a grappolo*, *stimatore* $\widehat{\mu}_{gr}$). Naturalmente, affinché un confronto di questo tipo abbia senso, deve essere effettuato a parità di unità elementari campionate. Detto n tale numero, si supporrà quindi che $n = m\,L$ (ovviamente è anche $N = M\,L$). È facile vedere, con ovvia simbologia, che vale la relazione:

$$\begin{aligned} \frac{V(\widehat{\mu}_{gr};\ grap)}{V(\overline{y}_{\mathbf{s}};\ ssr)} &\approx \frac{\left(\frac{1}{n} - \frac{1}{N}\right)\, S_y^2\,(1+(L-1)\,\rho_{ic})}{\left(\frac{1}{n} - \frac{1}{N}\right)\, S_y^2} \\ &= 1+(L-1)\,\rho_{ic}. \end{aligned} \tag{9.18}$$

L'esame della (9.18) permette di osservare che se il coefficiente di correlazione intra-classi ρ_{ic} è positivo, si ha (in via approssimata) $V(\overline{y}_{\mathbf{s}};\ ssr) < V(\widehat{\mu}_{gr};\ grap)$, e quindi l'uso del disegno semplice senza ripetizione fornisce (a parità di numerosità campionaria) risultati migliori rispetto a quello a grappolo. Se invece il coefficiente di correlazione intra-classi ρ_{ic} è negativo, il disegno a grappolo è da preferirsi a quello semplice.

A prima vista, questo risultato sembrerebbe sfavorire il disegno a grappolo, in quanto dalla relazione (9.14) appare chiaro che ben difficilmente ρ_{ic} assume

valori negativi. Tuttavia, è da rimarcare che spesso il disegno ssr è, a parità di numerosità campionaria, molto più costoso di quello a grappolo. Ciò è dovuto al fatto che molto spesso i grappoli sono composti da unità fisicamente "vicine", e quindi vi è un considerevole risparmio di costi nell'osservarle. A *parità di costo di rilevazione*, quindi, *il disegno a grappolo permette spesso di osservare più unità elementari rispetto al disegno ssr*, e questo potrebbe rendere la varianza $V(\widehat{\mu}_{gr};\ grap)$ più piccola di $V(\overline{y}_{\mathbf{s}};\ ssr)$. Per approfondimenti su questo punto si rinvia all'Esercizio 9.4.

Esempio 9.3. Consideriamo una scuola di $M = 100$ classi e supponiamo che ogni classe sia composta da $L = 25$ studenti. Il carattere di interesse è il numero di libri letto dagli studenti, di cui si vuole stimare la media sulla popolazione dei 2500 studenti. Supponiamo che $D_y^2 = 4410$ e $D_b^2 = 393$.

Supponiamo di stimare il numero medio di libri letti sulla base di un campione a grappolo composto da 15 classi. In ciascuna classe selezionata il numero medio di libri letto risulta pari alle seguenti quantità:

$$\begin{array}{cccccccc} 1.84 & 2.16 & 1.80 & 1.96 & 1.48 & 1.96 & 2.16 & 1.64 \\ 2.44 & 1.56 & 1.88 & 2.24 & 2.04 & 1.04 & 0.96. & \end{array}$$

Lo stimatore a grappolo assume di conseguenza il valore

$$\widehat{\mu}_{gr} = 1.82$$

e la sua varianza è pari a

$$V(\widehat{\mu}_{gr}) = 0.009\,.$$

Supponiamo inoltre di stimare il numero medio di libri letto sulla base di un campione casuale semplice senza ripetizione della stessa dimensione $n = mL = 375$. La media campionaria risulta pari a

$$\overline{y}_{\mathbf{s}} = 1.94$$

e la sua varianza è data da

$$V(\overline{y}_{\mathbf{s}}) = 0.004\,.$$

A parità di numerosità campionaria, in questo caso si ha che $V(\overline{y}_{\mathbf{s}};\ ssr) < V(\widehat{\mu}_{gr};\ grap)$. Notiamo che essendo il coefficiente di correlazione intra-classi pari a $\rho_{ic} = 0.05$, la correlazione positiva implica una perdita di precisione del campionamento a grappolo rispetto al campionamento casuale semplice. Formalmente, a parità di numerosità campionaria, da

$$\frac{V(\widehat{\mu}_{gr};\ grap)}{V(\overline{y}_{\mathbf{s}};\ ssr)} \approx 2$$

si desume che il campionamento casuale semplice risulta circa due volte più preciso del campionamento a grappolo. □

La (9.18) fornisce l'effetto del disegno nel caso di disegno a grappolo con grappoli di eguale ampiezza, e si può ovviamente riscrivere come:

$$Deff\,(grap,\ \overline{y}_{\mathbf{s}}) = 1 + (L-1)\rho_{ic} \tag{9.19}$$

essendo ρ_{ic} il coefficiente di correlazione intra-classi.

9.4 Grappoli di diversa numerosità e stima per quoziente

Le proprietà dello stimatore $\widehat{\mu}_{gr}$ (9.3) sono state studiate nella Sezione 9.2 in maniera del tutto generale, sia nel caso in cui i grappoli hanno la stessa numerosità, sia quando le loro numerosità sono differenti. Quest'ultimo caso merita però un esame più accurato.

Come si è già avuto modo di osservare, l'efficienza dello stimatore $\widehat{\mu}_{gr}$ è essenzialmente legata alla variabilità dei totali $T_1, \ldots, T_M$ dei grappoli. Quanto più bassa è tale variabilità, ossia quanto più "simili" sono i totali dei diversi grappoli, tanto più efficiente è lo stimatore $\widehat{\mu}_{gr}$.

Ora, l'esperienza pratica mostra che in parecchi casi di interesse i totali T_1, …, T_M dei grappoli esibiscono un'alta variabilità. Questo accade soprattutto nei casi in cui i grappoli hanno numerosità $N_1, \ldots, N_M$ molto differenti, mentre le loro medie $\mu_{y1}, \ldots, \mu_{yM}$ sono simili. Essendo i totali T_g pari a $N_g\,\mu_{yg}$, $g = 1, \ldots, M$, l'effetto finale sarà un'alta variabilità di $T_1, \ldots, T_M$.

Esempio 9.4. In una facoltà di Scienze Statistiche sono impartiti quattro corsi di Matematica 1, a classi rispettivamente di 10, 50, 25, 15 studenti. All'esame finale, tutti gli studenti conseguono lo stesso voto: 25.

Uno statistico vuole stimare il voto medio conseguito dagli studenti della facoltà, e decide di effettuare un campionamento a grappolo, in cui i grappoli sono le classi e il numero di grappoli campionati è pari a due. Come stimatore della media della popolazione decide poi di usare $\widehat{\mu}_{gr}$. Chiaramente, nel nostro caso è $N = 100$, $M = 4$, $m = 2$. Le medie dei grappoli, così come la media della popolazione, sono pari a 25. Le numerosità dei grappoli, i pesi, i totali e i valori z_g sono riportati in Tabella 9.2.

Tabella 9.2 Valori N_g, w_g, μ_{yg}, T_g

Grappolo g	*Numerosità* N_g	*Peso* w_g	*Media* μ_{yg}	*Totale* T_g	*Quantità* z_g
1	10	0.1	25	250	10
2	50	0.5	25	1250	50
3	25	0.25	25	625	25
4	15	0.15	25	275	15

In Tabella 9.3 sono invece enumerati tutti i possibili campioni di $m = 2$ grappoli, e per ciascuno di essi è calcolato il valore assunto dallo stimatore $\widehat{\mu}_{gr}$.

Tabella 9.3 Valori di $\widehat{\mu}_{gr}$ per campioni di $m = 2$ grappoli

Grappoli campionati	*Stima* $\widehat{\mu}_{gr}$
{1, 2}	$\frac{10+50}{2} = 30$
{1, 3}	$\frac{10+25}{2} = 17.5$
{1, 4}	$\frac{10+15}{2} = 12.5$
{2, 3}	$\frac{50+25}{2} = 37.5$
{2, 4}	$\frac{50+15}{2} = 32.5$
{3, 4}	$\frac{25+15}{2} = 20$

Il valore atteso di $\widehat{\mu}_{gr}$ è pari a $\mu_y = 25$, mentre la sua varianza è uguale a:

$$V(\widehat{\mu}_{gr}) = \left(\frac{1}{2} - \frac{1}{4}\right) \left\{\frac{1}{3}\sum_{g=1}^{4}(z_g - \mu_y)^2\right\} = 79.2.$$

La Tabella 9.3 mostra chiaramente come lo stimatore $\widehat{\mu}_{gr}$ fluttui molto attorno al suo valore atteso $\mu_y = 25$, in conseguenza dell'alta variabilità dei totali T_g dei grappoli. Per alcuni campioni $\widehat{\mu}_{gr}$ assume valori estremamente bassi, mentre per altri campioni $\widehat{\mu}_{gr}$ ha valori inaccettabilmente elevati. □

In situazioni di questo tipo lo stimatore $\widehat{\mu}_{gr}$ ha un'alta varianza, e quindi un'efficienza estremamente limitata. Ciò motiva la ricerca di qualche stimatore alternativo, che possa fornire risultati migliori quando le numerosità dei grappoli sono molto variabili.

9.4.1 Stimatore per quoziente

Come mostrato nell'Esempio 9.4, lo stimatore $\widehat{\mu}_{gr}$ è particolarmente inefficiente quando i totali dei grappoli sono molto variabili in conseguenza di un'alta variabilità delle loro numerosità, mentre le medie dei grappoli sono relativamente stabili. Ora, questo equivale a dire che i totali dei grappoli possono essere considerati, in via largamente approssimata, *proporzionali* alle relative numerosità:

$$T_g \approx cost\, N_g, \quad g = 1, \ldots, M. \tag{9.20}$$

Tenendo presenti i ragionamenti svolti nella Sezione 6.1, la (9.20) suggerisce di stimare μ_y tramite uno stimatore di tipo quoziente, in cui il ruolo della variabile ausiliaria $\mathcal{X}$ è svolto dalla numerosità dei grappoli:

$$x_g = N_g, \quad g = 1, \ldots, M.$$

Essendo

$$\mu_x = \frac{1}{M}\sum_{g=1} MN_g = \frac{N}{M}$$

si ha in tal modo lo stimatore di μ_y:

$$\begin{aligned}\widehat{\mu}_{qgr} &= \frac{\frac{1}{m}\sum_{g\in\mathbf{g}_m} z_g}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} x_g}\,\mu_x \\ &= \frac{\widehat{\mu}_{gr}}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} N_g}\,\frac{N}{M} \\ &= \frac{1}{M}\,\frac{\widehat{\mu}_{gr}}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} w_g} \\ &= \frac{1}{M}\,\frac{\widehat{\mu}_{gr}}{\overline{w}_m} \end{aligned} \tag{9.21}$$

dove si è posto

$$\overline{w}_m = \frac{1}{m}\sum_{g\in\mathbf{g}_m} w_g = \textit{media campionaria dei pesi dei grappoli.}$$

Esempio 9.5. Consideriamo ancora l'Esempio 9.4, e costruiamo lo stimatore $\widehat{\mu}_{qgr}$. In Tabella 9.4 sono enumerati tutti i possibili campioni di $m = 2$ grappoli; per ciascuno di essi sono calcolati i valori di $\widehat{\mu}_{gr}$, $\overline{w}_m$, e $\widehat{\mu}_{qgr}$.

Tabella 9.4 Valori di $\widehat{\mu}_{gr}$ per campioni di $m = 2$ grappoli

Grappoli campionati	*Stima* $\widehat{\mu}_{gr}$	$\overline{w}_m$	$\widehat{\mu}_{qgr}$
{1, 2}	$\frac{10+50}{2} = 30$	$\frac{0.1+0.5}{2} = 0.3$	$\frac{1}{4}\,\frac{30}{0.3} = 25$
{1, 3}	$\frac{10+25}{2} = 17.5$	$\frac{0.1+0.25}{2} = 0.175$	$\frac{1}{4}\,\frac{17.5}{0.175} = 25$
{1, 4}	$\frac{10+15}{2} = 12.5$	$\frac{0.1+0.15}{2} = 0.125$	$\frac{1}{4}\,\frac{12.5}{0.125} = 25$
{2, 3}	$\frac{50+25}{2} = 37.5$	$\frac{0.5+0.25}{2} = 0.375$	$\frac{1}{4}\,\frac{37.5}{0.375} = 25$
{2, 4}	$\frac{50+15}{2} = 32.5$	$\frac{0.5+0.15}{2} = 0.325$	$\frac{1}{4}\,\frac{32.5}{0.325} = 25$
{3, 4}	$\frac{25+15}{2} = 20$	$\frac{0.25+0.15}{2} = 0.2$	$\frac{1}{4}\,\frac{20}{0.2} = 25$

L'uso dello stimatore di tipo quoziente $\widehat{\mu}_{qgr}$ porta, nel presente esempio, a risultati nettamente migliori rispetto a quelli che si ottengono con $\widehat{\mu}_{gr}$. La ragione per cui questo accade è molto semplice. Nel nostro esempio i totali dei grappoli sono proporzionali alle numerosità dei grappoli stessi. In queste condizioni, mentre lo stimatore $\widehat{\mu}_{gr}$ fornisce risultati tanto peggiori quanto più

alta è la variabilità delle numerosità dei grappoli, lo stimatore $\widehat{\mu}_{qgr}$ fornisce risultati buoni, in quanto tale variabilità è controbilanciata dalla media campionaria $\overline{w}_m$ presente al denominatore. □

In via approssimata, il valore atteso e la varianza dello stimatore $\widehat{\mu}_{qgr}$ possono essere ricavati seguendo le stesse linee già sviluppate nella Sezione 6.2 per lo stimatore per quoziente. È sufficiente tenere conto che il disegno campionario è in effetti un disegno ssr sui grappoli, e che al posto dei valori y_i, x_i si hanno rispettivamente z_g e N_g, così che è:

$$R = \frac{\frac{1}{M}\sum_g z_g}{\frac{1}{M}\sum_g N_g} = \frac{M}{N}\,\mu_y. \tag{9.22}$$

I risultati della successiva Proposizione 9.5 sono un facile adattamento di quelli della Proposizione 6.1 e della (6.10).

Proposizione 9.5. *Se il disegno campionario è a grappolo, si ha:*

$$E\left[\widehat{\mu}_{qgr}\right] \approx \mu_y; \tag{9.23}$$

$$V(\widehat{\mu}_{qgr}) \approx \left(\frac{1}{m} - \frac{1}{M}\right)\left\{\frac{1}{M-1}\sum_{g=1}^{M}\left(z_g - \frac{M}{N}\,N_g\,\mu_y\right)^2\right\}. \tag{9.24}$$

Usando i risultati della Sezione 6.3 è anche facile, infine, costruire uno stimatore della varianza di $\widehat{\mu}_{qgr}$. Detta infatti

$$\begin{aligned}\widehat{R} &= \frac{\frac{1}{m}\sum_{g\in\mathbf{g}_m} z_g}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} N_g}\\ &= \frac{1}{N}\,\frac{\widehat{\mu}_{gr}}{\overline{w}_m}\end{aligned}$$

la "controparte campionaria" del rapporto R introdotto nella (9.22), è intuitivo fare riferimento a

$$\begin{aligned}\widehat{V}_{qgr} &= \left(\frac{1}{m} - \frac{1}{M}\right)\left\{\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}\left(z_g - \widehat{R}\,N_g\right)^2\right\}\\ &= \left(\frac{1}{m} - \frac{1}{M}\right)\left\{\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}\left(M\,w_g\,\mu_{yg} - \frac{\widehat{\mu}_{gr}}{\overline{w}_m}\,w_g\right)^2\right\}\end{aligned} \tag{9.25}$$

come stimatore di $V(\widehat{\mu}_{qgr})$ (9.24).

9.4.2 Considerazioni sull'efficienza dello stimatore per quoziente

L'espressione approssimata (9.24) permette di effettuare utili considerazioni sull'efficienza dello stimatore $\widehat{\mu}_{qgr}$. Infatti, una piccola rielaborazione della

(9.24) conduce alla seguente relazione:

$$V(\widehat{\mu}_{qgr}) \approx \left(\frac{1}{m} - \frac{1}{M}\right) \left\{ \frac{1}{M-1} \sum_{g=1}^{M} \left(M\, w_g\, \mu_{yg} - \frac{M}{N}\, N_g\, \mu_y \right)^2 \right\}$$

$$= \left(\frac{1}{m} - \frac{1}{M}\right) \left\{ \frac{1}{M-1} \sum_{g=1}^{M} \left(M\, w_g\, \mu_{yg} - M\, w_g\, \mu_y \right)^2 \right\}$$

$$= \left(\frac{1}{m} - \frac{1}{M}\right) \left\{ \frac{M^2}{M-1} \sum_{g=1}^{M} w_g^2\, \left(\mu_{yg} - \mu_y \right)^2 \right\}. \qquad (9.26)$$

Dalla (9.26) appare chiaro che la varianza di $\widehat{\mu}_{qgr}$ è tanto più piccola quanto più piccolo è il termine:

$$\frac{1}{M-1} \sum_{g=1}^{M} w_g^2\, \left(\mu_{yg} - \mu_y \right)^2$$

ovvero quanto più bassa è la variabilità delle medie μ_{yg} dei grappoli. Ciò significa, in sostanza, che lo stimatore quoziente $\widehat{\mu}_{qgr}$ è tanto più efficiente quanto più le medie dei grappoli tendono ad essere "simili" tra loro.

9.5 La progettazione di un disegno campionario a grappolo

Il campionamento a grappolo, come già si è avuto modo di osservare, è di frequente utilizzato, soprattutto perché spesso permette di osservare unità "vicine" nello spazio, con vantaggi notevoli di tempo e di costo di rilevazione. In effetti, accade spesso che i grappoli siano formati da unità elementari contigue, la cui osservazione è in genere molto più economica rispetto a quanto accade nel caso di disegno ssr, in cui le unità campionarie sono assai più "sparse".

9.5.1 Scelta della dimensione dei grappoli: qualche considerazione

A volte i grappoli sono suggeriti in modo naturale dall'oggetto della rilevazione. Altre volte, invece, il decidere di effettuare una rilevazione campionaria mediante disegno a grappolo implica che si devono risolvere due fondamentali problemi: (a) il numero M dei grappoli in cui suddividere la popolazione; (b) il numero di unità elementari da cui sono formati i grappoli. Nel seguito supporremo sempre che i grappoli abbiano tutti la *stessa numerosità* L, per cui i due problemi (a), (b) sono equivalenti (ogni grappolo contiene $L = N/M$ unità elementari).

Da un punto di vista intuitivo, quanto più elevato è il numero di unità che formano un grappolo, tanto maggiore è la variabilità del grappolo stesso. In altre parole, quanto più grande è L, tanto maggiore è la variabilità nei grappoli. Tenendo presente la relazione (9.13), e la simbologia usata nella Sezione 9.3, quanto detto equivale ad affermare che quanto più grande è L (e quindi quanto più piccolo è M), tanto più grande sarà il termine:

$$D_w^2 = \sum_g \sum_i (y_{gi} - \mu_{yg})^2$$

e tanto più piccolo sarà il termine

$$D_b^2 = L \sum_g (\mu_{yg} - \mu_y)^2.$$

Quanto finora detto, d'altra parte, significa che quanto più grande è L, tanto più piccolo tende ad essere il termine S_b^2 (9.4) che compare nella varianza dello stimatore $\widehat{\mu}_{gr}$. Da questo punto di vista è quindi vantaggioso scegliere un valore di L elevato, ossia pianificare grappoli formati da parecchie unità elementari. Tuttavia, questo si scontra con altre due considerazioni altrettanto importanti.

1. In grappoli "grandi" sono presenti anche unità elementari "lontane", e questo accresce il costo di osservazione per unità elementare.
2. Più grandi sono i grappoli, più piccolo è il numero di grappoli che, a parità di costo di rilevazione, entrano nel campione. In altre parole, quanto più grande è L, tanto più piccolo è m, e ciò tende a far crescere la varianza di $\widehat{\mu}_{gr}$.

In generale, scegliere la numerosità L dei grappoli, e quindi anche il numero $M = N/L$ dei grappoli da cui la popolazione è formata, è un compito difficile, che richiede la disponibilità di informazioni *a priori* sulla popolazione oggetto di studio, ed in particolare sulla variabilità del carattere di interesse. In linea di principio, l'idea è quella di usare le informazioni di cui si dispone per costruire:

1. una relazione che lega L e S_b^2;
2. una funzione di costo che lega L e m al *budget* disponibile per la rilevazione

e di usare le relazioni in 1, 2 per determinare il valore di L che, per un predeterminato costo di indagine, rende massima l'efficienza dello stimatore $\widehat{\mu}_{gr}$.

Tale approccio, benché molto intuitivo, richiede sia una notevole esperienza statistica, sia una grossa dose di informazioni empiriche, provenienti o da altre indagini su popolazioni "simili" a quella oggetto di studio, o da indagini effettuate nel passato sulla stessa popolazione. Un esempio di questo approccio è offerto nel volume di Cochran (1977), pp. 243–246.

9.5.2 Scelta del numero di grappoli del campione

Un secondo problema di considerevole importanza riguarda la scelta del numero m di grappoli da campionare.

La strada più semplice è quella di procedere in maniera simile a quanto già fatto per il disegno stratificato. L'obiettivo è quello di determinare la numerosità campionaria m in maniera tale che l'errore assoluto di stima $|\widehat{\mu}_{gr} - \mu_y|$ sia superiore ad una soglia t con probabilità pari a α, con t, α fissati a priori. In simboli:

$$Pr\left(|\widehat{\mu}_{gr} - \mu_y| > t\right) = \alpha. \tag{9.27}$$

La distribuzione di probabilità di $\widehat{\mu}_{gr}$ verrà approssimata con una normale di media μ_y e varianza (9.6). Ciò implica che la v.a. standardizzata

$$\frac{\widehat{\mu}_{gr} - \mu_y}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}}$$

ha in via approssimata distribuzione normale standard $N(0,\, 1)$. Ne consegue che la (9.27) si può rielaborare nel modo seguente

$$\begin{aligned} Pr\left(|\widehat{\mu}_{gr} - \mu_y| > t\right) &= Pr\left(\frac{|\widehat{\mu}_{gr} - \mu_y|}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}} > \frac{t}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}}\right) \\ &\approx Pr\left(|N(0,\, 1)| > \frac{t}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}}\right) \\ &= 2Pr\left(N(0,\, 1) > \frac{t}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}}\right) \\ &= \alpha \end{aligned}$$

da cui si ottiene la relazione

$$Pr\left(N(0,\, 1) > \frac{t}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}}\right) = \frac{\alpha}{2}. \tag{9.28}$$

Usando ragionamenti già visti nei capitoli precedenti, dalla (9.28) discende che

$$\frac{t}{\sqrt{\left(\frac{1}{m} - \frac{1}{M}\right) S_b^2}} = z_{\alpha/2}$$

da cui, con pochi passaggi, si ottiene la seguente espressione per il numero di grappoli del campione:

$$m = \frac{\frac{z_{\alpha/2}^2}{t^2} S_b^2}{1 + \frac{1}{M}\frac{z_{\alpha/2}^2}{t^2} S_b^2}. \tag{9.29}$$

Per M "grande" il termine $(z_{\alpha/2}/t)^2 S_b^2/M$ è sostanzialmente trascurabile, per cui la (9.29) si riduce a $m = (z_{\alpha/2}/t)^2 S_b^2$.

L'uso effettivo della (9.29) implica che si deve conoscere S_b^2, ossia la varianza tra i grappoli. In assenza di una tale informazione si può utilizzare una tecnica simile a quella del campione pilota descritta nella Sezione 4.3. L'idea di base è molto semplice. Si seleziona un campione iniziale (a grappolo) di numerosità abbastanza piccola, e con cui si stima S_b^2; tale stima verrà poi usata in (9.29) in luogo della "vera" S_b^2. In alternativa si possono usare, se disponibili, stime ottenute da rilevazioni precedenti sulla stessa popolazione, o su popolazioni simili.

Un metodo alternativo, simile nella sostanza a quello esposto nel Capitolo 7 per il disegno stratificato, è basato sull'effetto del disegno (vds. Capitolo 3). Per ragioni di semplicità ci si limiterà nel seguito al caso in cui le numerosità dei grappoli sono tutte uguali a L, così che $\widehat{\mu}_{gr}$ coincide con la media campionaria $\overline{y}_{\mathbf{s}}$.

Come già visto nella (9.19), si ha

$$Deff\,(grap,\ \overline{y}_{\mathbf{s}}) = \frac{V(\overline{y}_{\mathbf{s}};\ grap)}{V(\overline{y}_{\mathbf{s}};\ ssr)} \approx 1 + (L-1)\,\rho_{ic}.$$

Se, sulla base di precedenti rilevazioni o di un campione pilota, è noto *a priori* il valore di $Deff\,(grap,\ \overline{y}_{\mathbf{s}})$ (o almeno una sua stima sufficientemente accurata) ci si può basare su di esso per scegliere la numerosità campionaria n. Il procedimento consta di due fasi:

- fissati i valori di t e di α, si determina la numerosità campionaria n_{ssr} necessaria affinché sia $Pr(|\overline{y}_{\mathbf{s}} - \mu_y| > t) = \alpha$, secondo le linee esposte nel Capitolo 4;
- si calcola $m = \frac{n_{ssr}}{L}\, Deff\,(grap,\ \overline{y}_{\mathbf{s}})$, che fornisce il numero di grappoli campionari richiesto.

Esercizi

9.1. Sia $n = \sum_{g \in \mathbf{g}_m} N_g$ il numero di unità elementari osservate con un campionamento a grappolo. Provare che:

a. $E[n] = \frac{m}{M} N$

b. $V(n) = \left(\frac{1}{m} - \frac{1}{M}\right) \left\{ \frac{m^2}{M} \sum_{g=1}^{M} \left(N_g - \frac{N}{M}\right)^2 \right\}$.

Suggerimento. $n = \frac{1}{m} \sum_{g \in \mathbf{g}_m} m\, N_g$.

9.2. Provare che vale la relazione (9.15).

Suggerimento. Tenere conto che $D_b^2 = D_y^2 - D_w^2 = D_y^2 - \frac{L-1}{L}(1 - \rho_{ic}) D_y^2 = D_y^2 \frac{1}{L}\,(1 + (L-1)\rho_{ic})$, e che $S_b^2 = \frac{D_b^2}{L(M-1)}$.

9.3. Provare che il coefficiente di correlazione intra-classi si può esprimere nella forma:

$$\rho_{ic} = \frac{\frac{1}{M\,L\,(L-1)} \sum_g \sum_i \sum_{j \neq i} (y_{gi} - \mu_y)\,(y_{gj} - \mu_y)}{\frac{1}{M\,L} \sum_g \sum_i (y_{gi} - \mu_y)^2}\,.$$

Suggerimento. Tenere conto che $\sum_g (\mu_{yg} - \mu_y)^2 = \sum_g \{\sum_i (y_{gi} - \mu_y)\}^2 / L^2 = \{\sum_g \sum_i (y_{gi} - \mu_y)^2 + \sum_g \sum_i \sum_{j \neq i} (y_{gi} - \mu_y)\,(y_{gj} - \mu_y)\}/L^2$, da cui $D_w^2 = (L-1) D_y^2 / L - \sum_g \sum_i \sum_{j \neq i} (y_{gi} - \mu_y)\,(y_{gj} - \mu_y)/L$.

9.4. Si consideri una popolazione di N unità elementari, raggruppate in M grappoli ciascuno di numerosità L. Supponiamo poi che il costo di osservazione di un'unità elementare sia pari a c se si una il disegno ssr, e che sia pari a c/k, con $k \geqslant 1$, se si usa il disegno a grappolo. Fissato il *budget* totale C_0 per la rilevazione, si hanno due possibilità:

a. campionare mediante disegno ssr $n = C_0/c$ unità elementari;
b. campionare mediante disegno a grappolo $C_0/(k\,c) = k\,n$ unità elementari.

La *b.* richiede la selezione di $m = k\,n/L$ grappoli, per cui si suppone che tale numero sia intero.

Supponendo il numero totale N di unità "grande", così che $1/N \approx 0$, verificare che a parità di costo di rilevazione si ha

$$\frac{V(\overline{y}_{\mathbf{s}};\ ssr)}{V(\widehat{\mu}_{gr};\ grap)} \approx \frac{k}{1 + (L-1)\,\rho_{ic}}$$

così che la strategia *b.* è preferibile ad *a.* se $k > 1 + (L-1)\rho_{ic}$.

9.5. Una società di controllo di gestione deve valutare il numero medio di errori presenti in una serie di 10000 documenti contabili. Ogni documento contabile contiene 200 voci, ciascuna delle quali può o meno contenere un errore.

Si selezionano, con campionamento ssr, 100 documenti contabili, e si osserva che: (*i*) in 2 documenti ci sono 5 voci errate; (*ii*) in 3 documenti ci sono 2 voci errate; (*iii*) in 5 documenti vi è 1 voce errata; (*iv*) nei restanti 90 documenti non ci sono errori.

a. Considerando ogni documento contabile come un grappolo di 200 voci, stimare il numero medio di errori per documento contabile nella popolazione.
b. Stimare la varianza della stima in *a.*
c. Stimare il numero totale di voci errate nella popolazione di 10000 documenti contabili, e costruire una stima della varianza di tale stimatore.
d. Stimare il numero di documenti privi di errori nella popolazione.

9.6. Un naturalista vuole valutare il numero di piante di una data specie che vivono in un'area (quadrata) di 100 m^2. L'esame attento di 1 m^2 di superficie richiede 15 minuti di lavoro, e il naturalista può lavorare per non oltre 10 ore.

In via (molto) approssimata si può assumere che in una superficie di h m^2 ci si attendono in media $c\,h$ piante, con una varianza pari a $c\,h^r$, essendo r, c numeri reali positivi.

Il naturalista deve scegliere tra le seguenti due strategie di campionamento.

(i) Campionare mediante ssr 40 parcelle di 1 m^2 di terreno, contare il numero di piante che vivono in ciascuna parcella, e stimare il numero totale di piante come: 10000 × *media campionaria di piante per* m^2.

(ii) Suddividere il terreno in 2500 grappoli, ciascuno formato da 4 parcelle da 1 m^2, contare il numero di piante in ciascun grappolo, e stimare il numero totale di piante come: 10000 × *stimatore a grappolo del numero medio di piante per* m^2.

Provare che: (a) se $r = 1$ le due strategie sono equivalenti; se $r > 1$ è preferibile la strategia (i); se $r < 1$ è preferibile la strategia (ii).

Suggerimento. La varianza dello stimatore (i) è uguale, in via approssimata, a $\left(\frac{1}{40} - \frac{1}{10000}\right) \times c\,10000^2$. Sempre in via approssimata, la varianza dello stimatore (ii) è pari a $\left(\frac{1}{10} - \frac{1}{2500}\right) \times \frac{c\,4^r}{4^2}\,10000^2$.

9.7. Il sindaco di Frascati è in allarme, in quanto ha appreso che la vicina città di Roma vuole costruire, proprio vicino al confine tra i due comuni, un parcheggio per pullman turistici che inquinano e appesantiscono il traffico.

Per avere un'idea di cosa pensa la propria cittadinanza, il nostro sindaco decide di fare effettuare un'indagine campionaria. Viene selezionato un campione ssr di 100 delle 4000 famiglie di Frascati, e ad ogni famiglia viene chiesto di specificare: (a) il numero di adulti, e (b) il numero di adulti contrari alla costruzione del parcheggio. In totale, vengono in questo modo intervistate 260 persone adulte, 234 delle quali si dichiarano contrarie al parcheggio.

Sulla base dei risultati dell'indagine, il sindaco convoca una conferenza stampa e afferma che:

(i) sulla base di una seria indagine statistica, si è stimato che il 90% degli individui adulti di Frascati sono contrari alla costruzione del parcheggio;

(ii) la varianza della proporzione stimata in (i) si può stimare pari a $0.9\,(1 - 0.9)/260 = 0.00035$, e quindi l'indagine è molto affidabile.

Siete d'accordo, da un punto di vista statistico, con le affermazioni del sindaco?

9.8. In un quartiere vi sono 800 edifici, ciascuno identificato da via e numero civico. In ogni edificio vi è un certo numero di appartamenti (variabile da edificio a edificio), in ciascuno dei quali vive una famiglia. In totale il numero di famiglie che vivono nel quartiere è 16000. Si seleziona, mediante disegno ssr, un campione di $m = 20$ edifici di cui si riportano nella tabella che segue in numero di famiglie residenti e il reddito totale (somma dei redditi delle famiglie residenti).

Numero di famiglie	*Reddito totale (Euro)*	*Numero di famiglie*	*Reddito totale (Euro)*
18	35000	14	28000
24	50000	24	49000
26	51000	18	35000
15	31000	42	83000
30	63000	50	105000
32	62000	20	38000
36	75000	46	91000
25	49000	32	65000
18	38000	14	30000
20	43000	26	50000

a. Calcolare lo stimatore $\widehat{\mu}_{gr}$ del reddito medio per famiglia.
b. Stimare $V(\widehat{\mu}_{gr})$.
c. Calcolare $\widehat{\mu}_{qgr}$. Spiegare, sulla base dei dati campionari, perché $\widehat{\mu}_{qgr} < \widehat{\mu}_{gr}$.

Suggerimento. Che tipo di relazione suggeriscono i dati campionari?

10

Disegno campionario sistematico

10.1 Aspetti di base

Il disegno campionario sistematico è spesso usato in pratica, in quanto è in genere visto come un'alternativa semplificata del disegno ssr.

Per introdurre il campionamento sistematico, iniziamo con un semplice esempio.

Esempio 10.1. Si consideri una popolazione di $N = 20$ unità, da cui si vuole trarre un campione di $n = 5$ unità. Si può pensare di usare la seguente procedura di selezione.

Si considerano le prime 4 unità della popolazione, e si seleziona una di esse con probabilità 1/4:

- se si seleziona l'unità 1, il campione sarà $\mathbf{s}_1 = \{1, 5, 9, 13, 17\}$;
- se si seleziona l'unità 2, il campione sarà $\mathbf{s}_2 = \{2, 6, 10, 14, 18\}$;
- se si seleziona l'unità 3, il campione sarà $\mathbf{s}_3 = \{3, 7, 11, 15, 19\}$;
- se si seleziona l'unità 4, il campione sarà $\mathbf{s}_4 = \{4, 8, 12, 16, 20\}$.

Lo spazio dei campioni è pertanto composto dai seguenti quattro campioni:

$$\mathcal{S} = \{\mathbf{s}_1 = \{1, 5, 9, 13, 17\},\ \mathbf{s}_2 = \{2, 6, 10, 14, 18\},\ \mathbf{s}_3 = \{3, 7, 11, 15, 19\},\ \mathbf{s}_4 = \{4, 8, 12, 16, 20\}\}$$

e ciascuno di essi ha probabilità 1/4 di essere selezionato:

$$p(\mathbf{s}_1) = p(\mathbf{s}_2) = p(\mathbf{s}_3) = p(\mathbf{s}_4) = \frac{1}{4}.$$

Rispetto al disegno ssr vi sono differenze rilevanti, ma anche punti di contatto. La principale differenza è che nel disegno ssr lo spazio dei campioni è formato da *tutte* le combinazioni senza ripetizione di 5 delle 20 unità della popolazione. In questo caso, invece, sono presenti solo *alcune* di tali combinazioni. Il principale punto di contatto con il disegno ssr, invece, è costituito dal fatto che tutti i campioni hanno la stessa probabilità di essere selezionati. □

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_10,

In generale, consideriamo una popolazione di N unità, da cui si vuole selezionare un campione di numerosità n. Supponiamo anche, per semplicità, che $M = N/n$ sia intero (il caso N/n non intero sarà discusso più avanti); tale quantità è il *passo di campionamento*. Il *disegno campionario sistematico* di passo M è una facile estensione di quanto visto nell'Esempio 10.1, ed è di seguito descritto.

Si considerano le prime M unità della popolazione, e si seleziona una di esse con probabilità $1/M$:

- se si seleziona l'unità 1, il campione sarà $\mathbf{s}_1 = \{1,\ 1 + M,\ 1 + 2\,M,\ \ldots,\ 1 + (n-1)\,M\}$;
- se si seleziona l'unità 2, il campione sarà $\mathbf{s}_2 = \{2,\ 2 + M,\ 2 + 2\,M,\ \ldots,\ 2 + (n-1)\,M\}$;
  ...
- se si seleziona l'unità M, il campione sarà $\mathbf{s}_M = \{M,\ M + M,\ M + 2\,M,\ \ldots,\ M + (n-1)\,M\} = \{M,\ 2\,M,\ 3\,M,\ \ldots,\ N\}$.

Per quanto riguarda il disegno campionario, la sua struttura è chiara. Lo spazio dei campioni è formato dagli M campioni $\mathbf{s}_1,\ \mathbf{s}_2,\ \ldots,\ \mathbf{s}_M$:

$$\mathcal{S} = \{\{1,\ 1 + M,\ 1 + 2\,M,\ \ldots,\ 1 + (n-1)\,M\}\ ,\\ \{2,\ 2 + M,\ 2 + 2\,M,\ \ldots,\ 2 + (n-1)\,M\}, \ldots,\ \{M,\ 2\,M,\ 3\,M,\ \ldots,\ N\}\}$$

e ciascuno di essi ha probabilità $1/M$ di essere selezionato:

$$p(\mathbf{s}_1) = p(\mathbf{s}_2) = \ \ldots\ = p(\mathbf{s}_M) = \frac{1}{M}.$$

Differenze ed analogie con il disegno ssr sono evidenti. La differenza di base, in generale, è che i campioni selezionabili mediante disegno ssr sono tutte le $\binom{N}{n}$ combinazioni senza ripetizioni di n delle N unità della popolazione. Nel disegno sistematico, invece, solo $M = N/n$ di tali combinazioni sono effettivamente selezionabili. Questo significa che nel disegno ssr lo spazio dei campioni è più "ricco", composto da un maggior numero di campioni, rispetto al disegno sistematico. Inoltre, a differenza del campionamento casuale semplice nel campionamento sistematico solo la prima unità è scelta casualmente, mentre le altre sono determinate in modo automatico. Ciò semplifica notevolmente la procedura di estrazione del campione. La principale analogia tra i due disegni campionari, invece, è che in entrambi i casi tutti i campioni dello spazio dei campioni hanno la medesima probabilità di essere selezionati $1/\binom{N}{n}$ per il disegno ssr, $1/M = n/N$ per il disegno sistematico.

Il disegno sistematico è in sostanza *uno speciale tipo di disegno a grappolo*, come di seguito descritto.

- I grappoli sono gli insiemi di unità:
  $\{1, 1+M, 1+2\,M, \ldots, 1+(n-1)\,M\}$, $\{2, 2+M, 2+2\,M, \ldots, 2+(n-1)\,M\}$, ..., $\{M,\ 2\,M,\ 3\,M,\ \ldots,\ N\}$.

Questo significa che tutti i grappoli sono composti da n unità, e vi sono in totale $M = N/n$ grappoli. Con la simbologia introdotta nel precedente capitolo, questo significa che $L = n$.

– Si seleziona un solo grappolo: $m = 1$.

Quanto sopra detto vale soltanto nel caso in cui la numerosità N della popolazione sia un multiplo della numerosità n del campione, così che $M = N/n$ è un numero intero. Tuttavia, non sempre nella pratica tale condizione è soddisfatta. Per ovviare a tale inconveniente sono stati proposti diversi rimedi.

Una prima idea è quella di prendere M uguale al più piccolo intero maggiore o uguale a N/n, e di considerare M grappoli, di cui $M-1$ formati da n unità, e uno da $N - n\,(M-1)$ unità. Il modo in cui i grappoli sono costruiti e selezionati è del tutto simile a quello in precedenza descritto.

Si osservi che se N non è multiplo di n lo spazio dei campioni è costituito da campioni di numerosità diversa. Un modo per ovviare a tale inconveniente è quello di considerare la lista delle unità della popolazione come circolare al fine di selezionare un campione della numerosità prestabilita. Ciò significa che se si arriva alla fine della lista si riparte dall'inizio allo scopo di ottenere la numerosità campionaria prefissata. In questo caso si parla di *campionamento sistematico circolare*. Formalmente, tale metodo consiste nel prendere M uguale al più piccolo intero maggiore o uguale a N/n, e nell'"ampliare" la popolazione considerando $n\,M$ unità secondo lo schema seguente.

Unità della popolazione ampliata	*Unità della popolazione originale*
1	1
2	2
...	...
N	N
$N+1$	1
$N+2$	2
...	...
$n\,M$	$n\,M - N$

La popolazione ampliata viene poi suddivisa in M grappoli di n unità secondo il solito meccanismo, e viene selezionato un grappolo.

Esempio 10.2. Si consideri una popolazione di $N = 27$ unità, da cui si vuole trarre un campione sistematico di $n = 7$ unità. In questo caso si ha

$$M = \textit{più piccolo intero} \geqslant \frac{27}{7} = 4$$

e pertanto i grappoli in cui si suddivide la popolazione sono i seguenti:

$$\mathbf{s}_1 = \{1,\ 5,\ 9,\ 13,\ 17,\ 21,\ 25\},\ \mathbf{s}_2 = \{2,\ 6,\ 10,\ 14,\ 18,\ 22,\ 26\},$$
$$\mathbf{s}_3 = \{3,\ 7,\ 11,\ 15,\ 19,\ 23,\ 27\},\ \ \mathbf{s}_4 = \{4,\ 8,\ 12,\ 16,\ 20,\ 24\}.$$

Ognuno di essi, inoltre, ha probabilità 1/4 di essere selezionato. Di fatto, si procede come già visto, ossia:

Si seleziona una delle prime quattro unità, in modo che ciascuna abbia probabilità 1/4 di essere selezionata:

- se si seleziona l'unità 1, si osserva il campione $\mathbf{s}_1 = \{1, 5, 9, 13, 17, 21, 25\}$;
- se si seleziona l'unità 2, si osserva il campione $\mathbf{s}_2 = \{2, 6, 10, 14, 18, 22, 26\}$;
- se si seleziona l'unità 3, si osserva il campione $\mathbf{s}_3 = \{3, 7, 11, 15, 19, 23, 27\}$;
- se si seleziona l'unità 4, si osserva il campione $\mathbf{s}_4 = \{4, 8, 12, 16, 20, 24\}$.

□

L'inconveniente di questo metodo è che non produce campioni tutti della stessa numerosità n. Per particolari valori di N e n, uno dei campioni potrebbe essere composto da un numero molto piccolo di unità.

Esempio 10.3. Si consideri una popolazione di $N = 26$ unità da cui si vuole estrarre, con un disegno sistematico, un campione di $n = 7$ unità. Si ha in primo luogo

$$M = \textit{più piccolo intero} \geqslant \frac{26}{7} = 4,$$

per cui è $n\,M = 28$. La popolazione ampliata e quella originale sono riportate nello schema qui sotto.

Unità della popolazione ampliata	*Unità della popolazione originale*
1	1
2	2
...	...
26	26
27	1
28	2

A questo punto, come detto, si opera usando la popolazione ampliata come se fosse la popolazione da cui selezionare il campione sistematico. Questo significa procedere nel seguente modo.

Si seleziona una delle prime quattro unità (della popolazione ampliata), in modo che ciascuna abbia probabilità 1/4 di essere selezionata:

- se si seleziona l'unità 1, si osserva il campione (grappolo) di unità della popolazione ampliata {1, 5, 9, 13, 17, 21, 25}, che corrisponde al campione {1, 5, 9, 13, 17, 21, 25} della popolazione originale;
- se si seleziona l'unità 2, si osserva il campione (grappolo) di unità della popolazione ampliata {2, 6, 10, 14, 18, 22, 26}, che corrisponde al campione {2, 6, 10, 14, 18, 22, 26} della popolazione originale;
- se si seleziona l'unità 3, si osserva il campione (grappolo) di unità della popolazione ampliata {3, 7, 11, 15, 19, 23, 27}, che corrisponde al campione $\mathbf{s}_3 = \{3, 7, 11, 15, 19, 23, 1\}$ della popolazione originale;

– se si seleziona l'unità 4, si osserva il campione (grappolo) di unità della popolazione ampliata {4, 8, 12, 16, 20, 24, 28}, che corrisponde al campione {4, 8, 12, 16, 20, 24, 2} della popolazione originale. □

10.2 Stima della media della popolazione: risultati di base

Il problema della costruzione di uno stimatore della media della popolazione può essere trattato in maniera molto semplice tenendo conto che il disegno sistematico è un caso particolare di disegno a grappolo, in cui si seleziona un solo grappolo. Per ragioni di semplicità supporremo che $M = N/n$ sia intero. Molti dei risultati che si otterranno saranno comunque validi anche se si adottano gli schemi sistematici modificati (nel caso N/n non intero) visti nella sezione precedente. In ogni caso, per trattare il problema della stima della media della popolazione è sufficiente particolarizzare al caso $m = 1$ i risultati ottenuti nel capitolo precedente.

In primo luogo, gli $M = N/n$ grappoli in cui è suddivisa la popolazione possono essere indicati nel modo seguente.

Grappoli	*Unità dei grappoli*	*Medie dei grappoli*	*Valori* z_g
1	$\mathbf{s}_1 = \{1,\, 1+M,\, \ldots,\, 1+(n-1)\,M\}$	$\mu_{y1} = \frac{y_1+y_{1+M}+\cdots+y_{1+(n-1)\,M}}{n}$	$z_1 = \mu_{y1}$
2	$\mathbf{s}_2 = \{2,\, 2+M,\, \ldots,\, 2+(n-1)\,M\}$	$\mu_{y2} = \frac{y_2+y_{2+M}+\cdots+y_{2+(n-1)\,M}}{n}$	$z_2 = \mu_{y2}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
g	$\mathbf{s}_g = \{g,\, g+M,\, \ldots,\, g+(n-1)\,M\}$	$\mu_{y1} = \frac{y_g+y_{g+M}+\cdots+y_{g+(n-1)\,M}}{n}$	$z_g = \mu_{yg}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
M	$\mathbf{s}_M = \{M, 2\,M,\, \ldots,\, N\}$	$\mu_{yM} = \frac{y_M+y_{2\,M}+\cdots+y_N}{n}$	$z_1 = \mu_{yM}$

Se si seleziona il grappolo g-mo, lo stimatore $\widehat{\mu}_{gr}$ si riduce a:

$$\begin{aligned}\widehat{\mu}_{gr} &= \mu_{yg} \\ &= \frac{\textit{Somma delle } y_i \textit{ con } i \in \mathbf{s}_g}{n} \\ &= \textit{Media campionaria delle } y_i \\ &= \overline{y}_{\mathbf{s}}\end{aligned} \tag{10.1}$$

ossia alla *media campionaria*, per la quale useremo il simbolo $\overline{y}_{\mathbf{s}}$ già impiegato nel Capitolo 3.

Le proprietà dello stimatore (10.1) sono studiate nella Proposizione 10.1, che è semplicemente ottenuta particolarizzando la Proposizione 9.1 al caso $m = 1$. Per rendere più comoda la notazione, indichiamo con

$$S_b^2 = \frac{1}{M-1}\sum_{g=1}^{M}(\mu_{yg} - \mu_y)^2 \tag{10.2}$$

la varianza (corretta con il denominatore $M-1$) delle medie μ_{yg} nella popolazione.

Proposizione 10.1. *Se il disegno campionario è sistematico, la media campionaria* $\overline{y}_{\mathbf{s}}$ *è uno stimatore corretto della media della popolazione:*

$$E\left[\overline{y}_{\mathbf{s}}\right] = \mu_y \tag{10.3}$$

e la sua varianza è pari a

$$V\left(\overline{y}_{\mathbf{s}}\right) = \left(1 - \frac{n}{N}\right) S_b^2. \tag{10.4}$$

La varianza (10.4) si può studiare un po' più in dettaglio tramite il coefficiente di correlazione intra-classi ρ_{ic} introdotto nella Sezione 9.3. Esattamente come nella Sezione 9.3, consideriamo le quantità

$$D_y^2 = \sum_{i=1}^{N} (y_i - u_y)^2, \quad D_b^2 = n \sum_{g=1}^{M} (\mu_{yg} - \mu_y)^2$$

e

$$D_w^2 = \sum_{g=1}^{M} \left\{ (y_g - \mu_{yg})^2 + (y_{g+M} - \mu_{yg})^2 + \cdots + (y_{g+(n-1)\,M} - \mu_{yg})^2 \right\}$$

ovvero, rispettivamente, la *devianza totale*, la *devianza tra i grappoli* e la *devianza nei grappoli*. Il coefficiente di correlazione intra-classi è pari a:

$$\rho_{ic} = 1 - \frac{n}{n-1} \frac{D_w^2}{D_y^2}.$$

Indicando come al solito con $S_y^2 = D_y^2/(N-1)$ la varianza corretta della popolazione, dalla Proposizione 9.4 discende che vale la relazione:

$$V(\overline{y}_{\mathbf{s}}) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N-1}{N-n} S_y^2 \left(1 + (n-1)\,\rho_{ic}\right). \tag{10.5}$$

Se N è grande rispetto a n si ha $N - n \approx N - 1$, e quindi la (10.5) può essere approssimata nel modo seguente:

$$V(\overline{y}_{\mathbf{s}}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \left(1 + (n-1)\,\rho_{ic}\right). \tag{10.6}$$

La (10.6) permette di confrontare facilmente l'efficienza della coppia (*disegno ssr*, *media campionaria*) con quella della coppia (*disegno sistematico*, *media campionaria*), a parità di numerosità campionaria n. Si ha la relazione:

$$\frac{V(\overline{y}_{\mathbf{s}};\ sist)}{V(\overline{y}_{\mathbf{s}};\ ssr)} \approx 1 + (n-1)\,\rho_{ic} \tag{10.7}$$

la quale si può scrivere in termini di effetto del disegno come:

$$Deff\,(sist,\ \overline{y}_{\mathbf{s}}) \approx 1 + (n-1)\,\rho_{ic}\,. \tag{10.8}$$

Dalla (10.7) si evince che se il coefficiente di correlazione intra-classi ρ_{ic} è negativo, si ha (in via approssimata) $V(\overline{y}_{\mathbf{s}};\ sist) < V(\overline{y}_{\mathbf{s}};\ ssr)$, e quindi l'uso del disegno sistematico garantisce (in via approssimata) un'efficienza di stima superiore rispetto a quella ottenibile con il disegno semplice senza ripetizione. La disuguaglianza contraria vale invece se ρ_{ic} assume un valore positivo. In queste condizioni, notiamo che anche un piccolo valore del coefficiente di correlazione intra-classi può provocare un forte incremento nella varianza (10.6) a causa del fattore $(n-1)$, con una conseguente minore efficienza del campionamento sistematico rispetto al campionamento casuale semplice senza ripetizione.

Esempio 10.4. Nel *file* `agenzie02.txt` sono riportati il valore delle vendite e l'utile lordo (entrambi in milioni di Euro) di 100 agenzie immobiliari che operano nella provincia di Roma. Si tratta di una popolazione di $N = 100$ unità, di cui si vuole stimare il valore medio delle vendite, μ_y. In particolare, nella popolazione è $\mu_y = 16.6$, $S_y^2 = 80.6$.

Se si usa un disegno ssr con numerosità $n = 20$, la varianza e il coefficiente di variazione della media campionaria $\overline{y}_{\mathbf{s}}$ sono rispettivamente uguali a:

$$V(\overline{y}_{\mathbf{s}};\ ssr) = \left(\frac{1}{20} - \frac{1}{100}\right) 80.6 = 3.2;$$
$$CV(\overline{y}_{\mathbf{s}};\ ssr) = \frac{3.2}{16.6}\,100 = 19.3\%.$$

Confrontiamo questo risultato con quello che si avrebbe usando un disegno sistematico, sempre di numerosità $n = 20$. In totale si hanno $M = 100/20 = 5$ grappoli (campioni), le caratteristiche dei quali sono qui di seguito elencate.

Le devianze totale (D_y^2), nei grappoli (D_w^2) e tra i grappoli (D_b^2) sono qui sotto riportate

$$D_y^2 = 7978.5,\ \ D_w^2 = 7862.5,\ \ D_b^2 = 116.0.$$

Tabella 10.1 Caratteristiche dei grappoli per la popolazione di agenzie immobiliari

Grappolo	*Unità*	*Media*	*Devianza*
1	{1, 6, 11, ,$\cdots$, 96}	15.1	1003.9
2	{2, 7, 12, ,$\cdots$, 97}	18.2	1970.2
3	{3, 8, 13, ,$\cdots$, 98}	16.5	1973.2
4	{4, 9, 14, ,$\cdots$, 99}	15.9	1422.6
5	{5, 10, 15, ,$\cdots$, 100}	17.3	1492.6

Esse danno luogo ad un coefficiente di correlazione intra-classi pari a

$$\rho_{ic} = 1 - \frac{20}{19}\frac{7862.5}{7978.5} = -0.04.$$

Essendo il valore di ρ_{ic} negativo, si può affermare che il disegno sistematico dà risultati migliori, in termini di efficienza, rispetto al disegno ssr. Questo è confermato anche dal calcolo della varianza e del coefficiente di variazione di $\overline{y}_{\mathbf{s}}$

$$V(\overline{y}_{\mathbf{s}};\ sist) = \left(\frac{1}{20} - \frac{1}{100}\right)\frac{99}{80}80.6\,(1 - 19 \times 0.04) = 0.96;$$
$$CV(\overline{y}_{\mathbf{s}};\ sist) = \frac{0.96}{16.6}100 = 5.8\%$$

i quali mostrano che in questo caso il disegno sistematico comporta un notevole guadagno di efficienza rispetto al disegno ssr. □

10.3 Efficienza di stima con disegno sistematico

Come già sottolineato, il disegno sistematico può essere considerato come un caso particolare di quello a grappolo. È quindi soggetto, in linea di principio, a considerazioni simili per quanto riguarda l'efficienza di stima della media della popolazione. Vi sono però diverse peculiarità che vale la pena sottolineare.

I grappoli (campioni) propri del campionamento sistematico non sono, in genere, gruppi di unità legate da un qualche vincolo di contiguità spaziale o di altra natura. La loro natura dipende essenzialmente dal modo in cui le etichette sono assegnate alle unità della popolazione. Modi diversi di assegnare le etichette alle unità della medesima popolazione possono portare a valori completamente diversi della varianza $V(\overline{y}_{\mathbf{s}};\ sist)$. Pertanto, l'efficienza di stima che si ottiene usando un disegno di tipo sistematico dipende strettamente dal modo in cui le etichette sono assegnate alle unità della popolazione, ossia, in termini equivalenti, al modo in cui le unità sono ordinate prima di procedere all'estrazione del campione. Questo punto è illustrato nel successivo Esempio 10.5.

Esempio 10.5. Nel *file* `spese_anziani.xls` sono riportate diverse variabili relative a 250 comuni (i cui nomi sono di fantasia). Ciascun comune è identificato da un codice di tre cifre. I comuni sono raggruppati in 12 distretti (anch'essi con nomi di fantasia, e con codice numerico da 1 a 12), i quali a loro volta formano tre regioni (sempre con nomi di fantasia, e con codice numerico da 1 a 3). La variabile di interesse, di cui si vuole stimare la media, è la spesa (media) per anziano sostenuta nell'anno 2011. Per ciascun comune sono note *a priori* la popolazione residente e la spesa per anziani sostenuta nell'anno 2009.

Per stimare la spesa media per anziani sostenuta nel 2011, consideriamo in primo luogo un disegno campionario sistematico, di numerosità $n = 25$. Vi sono quindi, in totale, $M = 10$ grappoli, ciascuno dei quali è uno dei possibili campioni selezionabili. Per quanto riguarda l'assegnazione delle etichette alle unità, sono stati considerati quattro diversi criteri di ordinamento:

- *nome del comune*: i comuni sono ordinati alfabeticamente sulla base del proprio nome, e le etichette sono assegnate di conseguenza (al primo comune della graduatoria alfabetica viene assegnata l'etichetta 1, al secondo l'etichetta 2, e così via);
- *codice del comune*: i comuni sono ordinati sulla base del proprio codice numerico, e le etichette sono assegnate di conseguenza;
- *popolazione del comune*: i comuni sono ordinati sulla base della popolazione residente, e le etichette sono assegnate di conseguenza;
- *spesa per anziani nel* 2009: i comuni sono ordinati sulla base della propria spesa (media) per anziano sostenuta nel 2009, e le etichette sono assegnate di conseguenza.

Intuitivamente, l'usare come criterio di ordinamento il nome del comune è pressoché equivalente ad assegnare casualmente le etichette ai comuni. Considerazioni abbastanza simili si possono fare per l'ordinamento in base ai codici dei comuni, anche se in genere codici "simili" indicano la vicinanza geografica dei comuni corrispondenti. Poiché è ragionevole pensare che le spese sostenute nel 2011 siano (molto) correlate positivamente con quelle del 2009, è lecito ritenere che l'ordinare i comuni sulla base della spesa per anziano del 2009 dovrebbe grosso modo (anche se non esattamente) portare ad assegnare le etichette più piccole ai comuni con spesa (nel 2011) più bassa, e le etichette più grandi ai comuni con spesa (sempre nel 2011) più alta. Infine, l'ordinare i comuni sulla base della popolazione porta a risultati non facilmente prevedibili, in quanto legati alle relazioni tra spesa per anziano nel 2011 e livello di popolazione.

In Fig. 10.1 sono rappresentati grafici in cui in ascissa compaiono le etichette dei comuni, e in ordinata la spesa (media) per anziano nel 2011.

I grafici sono molto diversi, ma mostrano tutti un aspetto interessante. Molto grossolanamente, i comuni possono dividersi in due categorie: quella dei comuni di spesa medio-bassa (all'incirca tra i 5000 e i 6000 Euro annui per anziano), e quelli di spesa medio-alta (all'incirca tra i 15000 e i 17000 Euro per anziano). Inoltre, è chiara l'alta correlazione positiva tra spesa per anziano nel 2009 e nel 2011.

In Tabella 10.2 sono riportate le medie campionarie per i 10 campioni (di $n = 25$ comuni) del disegno sistematico, con i quattro criteri di ordinamento sopra elencati. Nella stessa tabella sono anche riportati devianza totale, devianze nei grappoli e tra i grappoli, e coefficiente di correlazione intra-classi, che nel caso in esame ha come valore minimo possible $-1/24 = -0.042$, e come valore massimo 1. Nella stessa tabella sono anche riportati valore atteso, varianza e deviazione standard della media campionaria sia quando il

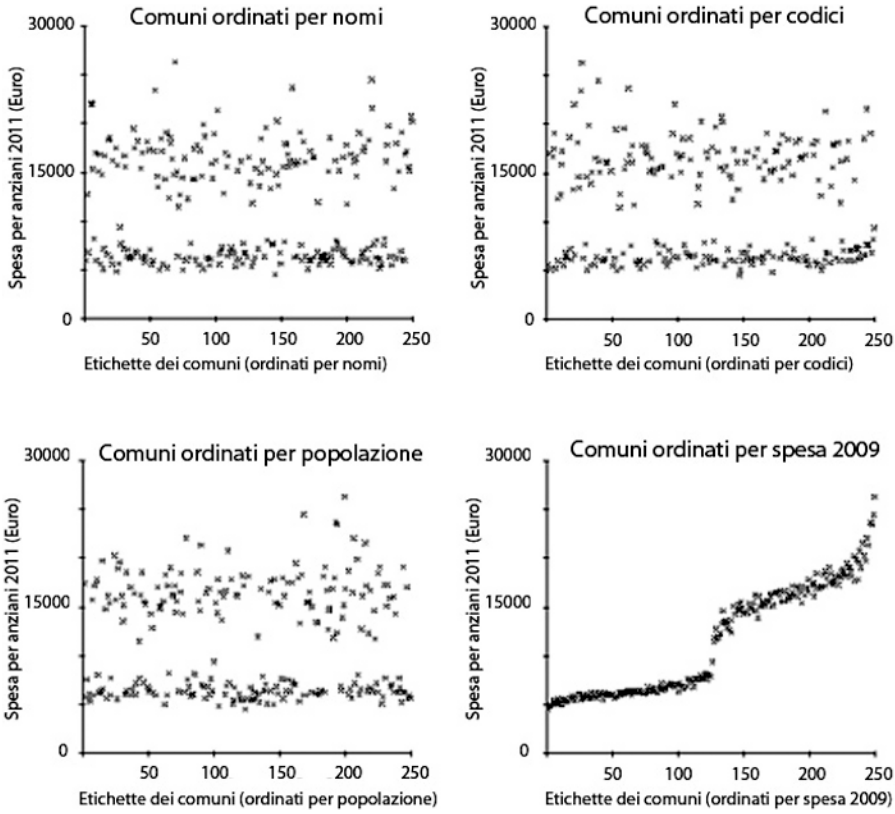


Fig. 10.1 Grafici spesa media per anziano del 2011

disegno è sistematico, sia quando è ssr. In quest'ultimo caso, ovviamente, la varianza della media campionaria non dipende dal modo in cui le etichette sono assegnate ai comuni.

Il disegno sistematico consente un guadagno di efficienza molto modesto rispetto al disegno ssr quando i comuni sono ordinati per nome, e un pò più marcato quando sono ordinati per codice. Questo è conforme all'intuizione che assegnare le etichette sulla base dei nomi è molto vicino ad una assegnazione puramente casuale, per cui i risultati sono tutto sommato comparabili (lievemente migliori nel nostro caso) a quelli che si ottengono con il disegno ssr. L'ordinamento dei comuni sulla base della popolazione produce invece un'efficienza minore di quella che si ha con il disegno ssr. Infine, ordinare i comuni sulla base della spesa per anziano del 2009 produce invece un notevolissimo guadagno di efficienza, come si vede sia dall'esame delle deviazioni standard, sia dal coefficiente di correlazione intra-classi (pari a -0.039).

È interessante cercare di capire a cosa sia dovuto un tale guadagno di efficienza. Ordinare i comuni sulla base della spesa per anziano del 2009 è

Tabella 10.2 Confronto disegni sistematico e ssr – stima spesa media per anziani

Campioni	*Disegno sistematico*			
	Ordinamento per			
	Nome comune	*Codice comune*	*Popolazione*	*Spesa 2009*
1	11682	10994	10003	11064
2	10815	11453	10912	11113
3	10375	12097	13137	11186
4	11049	10603	10113	11447
5	10460	12856	11177	11372
6	11776	11954	11725	11517
7	12044	11745	13615	11695
8	13639	10556	9865	11826
9	12298	11346	13022	11782
10	10835	11369	11404	11971
Dev. totale D_y^2	7583897917	7583897917	7583897917	7583897917
Dev. tra grappoli D_b^2	226984103	111402003	420433053	22588403
Dev. nei grappoli D_w^2	7356913814	7472495914	7163464864	7561309514
Coeff. corr. intra-classi ρ_{ic}	–0,01	–0,026	0,016	–0,039
Disegno sistematico – stimatore media campionaria				
Valore atteso	11497	11497	11497	11497
Varianza	907936	445608	1681732	90354
Deviazione standard	953	668	1297	301
Disegno ssr – stimatore media campionaria				
Valore atteso	11497			
Varianza	1096467			
Deviazione standard	1047			

quasi equivalente ad ordinarli sulla base della spesa per anziano del 2011. Ogni campione di 25 comuni, a causa di questo tipo di ordinamento, conterrà sia comuni con un livello di spesa basso/medio, sia comuni con un livello di spesa medio/alto. In altri termini, i campioni sono "ben equilibrati" nella composizione, e la media di ciascuno di essi è vicina a quella di tutti i 250 comuni. Lo stesso non necessariamente accade nel caso del disegno semplice. Queste considerazioni sono illustrate graficamente in Fig. 10.2. □

Dall'esempio precedente possono trarsi alcune considerazione sull'efficienza di stima che può ottenere con il disegno sistematico. Se i valori della variabile oggetto di indagine sono ordinati casualmente allora il campionamento sistematico equivale, in termini di efficienza, ad un campionamento casuale semplice. Per la dimostrazione formale, si veda l'Esercizio 10.1.

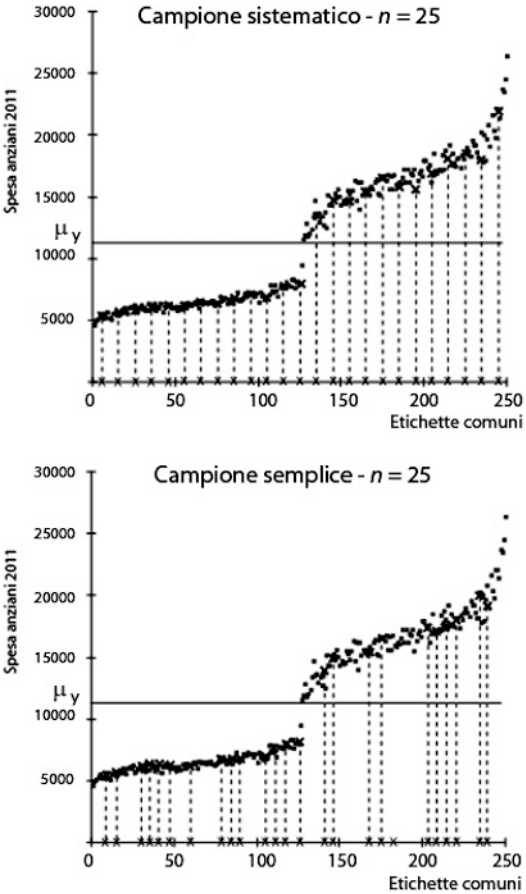


Fig. 10.2 Campioni sistematico e ssr – esempio spesa anziani 2011

Se invece i dati della variabile oggetto di indagine sono disposti in ordine crescente o decrescente, allora il campionamento sistematico può avere un'efficienza di stima maggiore di quella del campionamento casuale semplice. In particolare, se l'eterogeneità della popolazione viene colta dal passo di campionamento generando un campione costituito da elementi molto diversi tra loro (alcuni con valori y "piccoli" e altri con valori y "grandi"), il campionamento sistematico risulta migliore del campionamento casuale semplice. Sotto tale condizione il campione rappresenta in qualche misura un'"immagine ridotta" della popolazione, poiché sarà costituito da unità in cui la variabile d'indagine assume valori piccoli, medi ed elevati in proporzione simile alla popolazione. Si veda in proposito la Fig. 10.3.

Il campionamento sistematico può d'altro canto fornire risultati peggiori del campionamento casuale semplice se i valori della variabile di interesse presentano un andamento ciclico. In questa circostanza, se la ciclicità viene colta dal passo di campionamento il campione sarà costituito da elementi

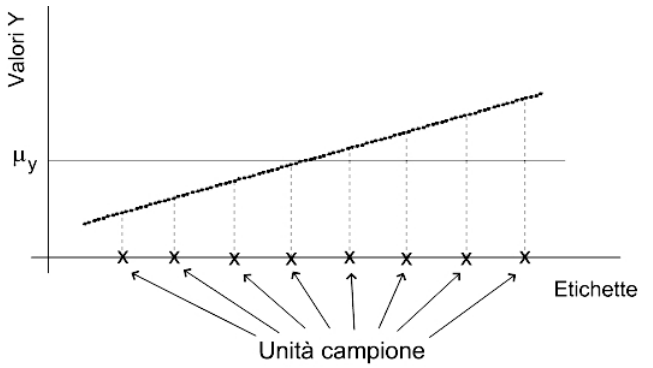


Fig. 10.3 Campione sistematico con valori y crescenti

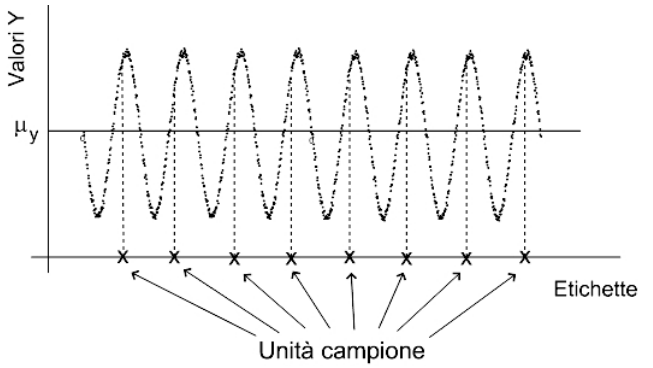


Fig. 10.4 Campione sistematico con valori y ciclici

molto simili tra loro, come appare in Fig. 10.4. In generale per popolazioni caratterizzate da periodicità il passo di campionamento non deve coincidere con il periodo o con un suo multiplo.

Esempio 10.6 (Trend lineare). Un caso speciale di popolazione con modalità y_i ordinate è quello di *trend lineare*, in cui si assume che i valori della variabile $\mathcal{Y}$ siano legati alle etichette delle unità corrispondenti dalla seguente relazione lineare

$$y_i = a + bi \quad \textit{per ciascuna unità } i = 1, \ldots, N \tag{10.9}$$

con a, b numeri reali. La media μ_y della popolazione è pari a

$$\begin{aligned}\mu_y &= \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(a+bi) = a + \frac{b}{N}\sum_{i=1}^{N} i = a + \frac{b}{N}\frac{N(N+1)}{2} \\ &= a + b\frac{N+1}{2}\end{aligned} \tag{10.10}$$

in quanto (Esercizio 10.2) la somma dei primi N numeri interi è pari a $N(N+1)/2$. In modo simile, è facile vedere che la varianza della popolazione assume la forma:

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(a+bi-a-b\frac{N+1}{2}\right)^2 \\
&= \frac{b^2}{N}\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 \\
&= b^2\left\{\frac{1}{N}\sum_{i=1}^{N} i^2 - \left(\frac{N+1}{2}\right)^2\right\} \\
&= b^2\left(\frac{1}{N}\frac{N(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4}\right) \\
&= b^2(N+1)\left\{\frac{2N+1}{6} - \frac{N+1}{4}\right\} \\
&= b^2(N+1)\frac{N-1}{12} \\
&= b^2\frac{N^2-1}{12}
\end{aligned}$$

essendo la somma dei quadrati dei primi N numeri naturali eguale a $N(N+1)(2N+1)/6$ (Esercizio 10.3). La varianza *corretta* della popolazione è pertanto pari a

$$\begin{aligned}
S_y^2 &= \frac{N}{N-1}\sigma_y^2 \\
&= \frac{N}{N-1}b^2\frac{N^2-1}{12} \\
&= b^2\frac{N(N+1)}{12}.
\end{aligned} \tag{10.11}$$

Se si seleziona dalla popolazione un campione semplice senza ripetizione di numerosità n, la varianza della media campionaria risulta uguale, con ovvia simbologia, a:

$$V(\overline{y}_{\mathbf{s}};\, ssr) = \left(\frac{1}{n}-\frac{1}{N}\right)b^2\frac{N(N+1)}{12}. \tag{10.12}$$

Supponiamo ora che $M = N/n$ sia intero, e consideriamo un disegno sistematico di ampiezza n (e passo M). Gli M campioni del disegno sistematico possono essere scritti come

$$\mathbf{s}_i = \{i,\, i+M,\, i+2M,\, \ldots,\, i+(n-1)M\},\quad i = 1,\, \ldots,\, M \tag{10.13}$$

e ad essi corrispondono le modalità etichettate

$$a + bi,\ a + b(i + M),\ a + b(i + 2M),\ \ldots, a + b(i + (n-1)M)\,. \quad (10.14)$$

Dalla (10.14) discende che al campione $\mathbf{s}_i$ corrisponde una media campionaria

$$\begin{aligned}
\overline{y}_{\mathbf{s}_i} &= \frac{a + bi + a + b(i + M) + a + b(i + 2M) + \ldots + a + b(i + (n-1)M)}{n} \\
&= a + \frac{b}{n}\left(ni + M\sum_{j=0}^{n-1} j\right) \\
&= a + bi + \frac{bM}{n}\sum_{j=1}^{n-1} j \\
&= a + bi + \frac{bM}{n}\frac{n(n-1)}{2} \\
&= a + bi + \frac{(n-1)bM}{2} \qquad (10.15)
\end{aligned}$$

sempre come conseguenza dell'Esercizio 10.2.

La varianza della media campionaria, se il disegno è sistematico, diventa

$$\begin{aligned}
V(\overline{y}_{\mathbf{s}};\ sist) &= \frac{1}{M}\sum_{i=1}^{M}(\overline{y}_{\mathbf{s}_i} - \mu_y)^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left\{a + bi + \frac{(n-1)bM}{2} - a - b\left(\frac{N+1}{2}\right)\right\}^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left(a + bi + b\frac{nM}{2} - b\frac{M}{2} - a - b\frac{N}{2} - \frac{b}{2}\right)^2 \\
&= \frac{b^2}{M}\sum_{i=1}^{M}\left(i - \frac{M+1}{2}\right)^2 \\
&= b^2\left\{\frac{1}{M}\sum_{i=1}^{M} i^2 - \left(\frac{M+1}{2}\right)^2\right\} \\
&= b^2\left\{\frac{1}{M}\frac{M(M+1)(2M+1)}{6} - \frac{(M+1)^2}{4}\right\} \\
&= b^2\frac{M^2-1}{12}
\end{aligned}$$

avendo ancora sfruttato l'Esercizio 10.3.

Il rapporto delle due varianze (10.16), (10.12) risulta pertanto pari a

$$\frac{V(\overline{y}_{\mathbf{s}};\ sist)}{V(\overline{y}_{\mathbf{s}};\ ssr)} = \frac{M+1}{N+1} < 1 \qquad (10.16)$$

e quindi si conclude che, a parità di numerosità campionaria n, il disegno sistematico fornisce nel caso in esame un'efficienza di stima maggiore rispetto al disegno semplice. □

In conclusione i vantaggi del campionamento sistematico sono essenzialmente di due tipi.

1. Se i valori della variabile di interesse $\mathcal{Y}$ sono disposti casualmente, il campionamento sistematico è assimilabile in tutto e per tutto a quello casuale semplice. Rispetto al campionamento casuale semplice è più facile da implementare poiché richiede l'utilizzazione di un meccanismo casuale soltanto per la selezione della prima unità.
2. È più efficiente di quello casuale semplice se le unità della popolazione possono essere ordinate secondo i valori di una variabile ausiliaria che risulta correlata con la variabile di interesse, come illustrato in precedenza.

10.4 Stima della varianza della media campionaria

Il campionamento sistematico, benché intuitivo e facile da eseguire, presenta uno svantaggio assai rilevante rispetto al disegno semplice senza ripetizione. Poiché si basa sull'osservazione di uno solo dei grappoli in cui è suddivisa la popolazione ($m = 1$), esso non permette di costruire uno stimatore corretto della varianza $V(\overline{y}_{\mathbf{s}})$.

Se le etichette sono assegnate alle unità in modo casuale, il disegno sistematico (Esercizio 10.1) è sostanzialmente equivalente al disegno semplice senza ripetizione. Una pratica comune, in tal caso, consiste nello stimare $V(\overline{y}_{\mathbf{s}})$ esattamente come visto per il disegno ssr, ossia tramite lo stimatore:

$$\widehat{V} = \left(\frac{1}{n} - \frac{1}{N}\right) \widehat{s}_y^2 \tag{10.17}$$

essendo

$$\widehat{s}_y^2 = \frac{1}{n-1} \sum_{i \in \mathbf{s}} (y_i - \overline{y}_{\mathbf{s}})^2$$

la varianza campionaria corretta. In maniera del tutto analoga vengono anche costruiti intervalli di confidenza per μ_y

In generale, lo stimatore (10.17) è distorto se le etichette non sono assegnate casualmente alle unità. Differenti stimatori della varianza (10.4) sono studiati nel volume di Wolter (2007); ad eccezione di casi particolari riguardanti la struttura della popolazione, si tratta comunque di stimatori distorti. Dato il livello elementare della presente trattazione non proseguiamo ulteriormente in questa direzione, limitandoci a presentare uno stimatore della (10.4) nell'Esercizio 10.4.

Esempio 10.7. Si consideri ancora la popolazione con trend lineare dell'Esempio 10.6, in cui si assume che $y_i = a + bi$, $i = 1, \ldots, N$. Per il campione $\mathbf{s}_i = \{i, i+M, i+2M, \ldots, i+(n-1)M\}$, alle cui unità corrispondono valori y_i del $a + b(i + jM)$, $j = 0, 1, \ldots, n-1$, la media campionaria è pari alla (10.15). La varianza campionaria corretta assume pertanto la forma:

$$\begin{aligned}
\widehat{s}_i^2 &= \frac{1}{n-1}\left\{(y_i - \overline{y}_s)^2 + (y_{i+M} - \overline{y}_s)^2 + \cdots + (y_{i+(n-1)M} - \overline{y}_s)^2\right\} \\
&= \frac{1}{n-1}\sum_{j=0}^{n-1}(y_{i+jM} - \overline{y}_s)^2 \\
&= \frac{1}{n-1}\sum_{j=0}^{n-1}\left\{a + b(i+jM) - \left(a + bi + \frac{(n-1)bM}{2}\right)\right\}^2 \\
&= \frac{b^2M^2}{n-1}\sum_{j=0}^{n-1}\left(j - \frac{n-1}{2}\right)^2 \\
&= \frac{b^2M^2}{n-1}\left(\sum_{j=0}^{n-1} j^2 - n\frac{(n-1)^2}{4}\right) \\
&= \frac{b^2M^2}{n-1}\left(\frac{n(n-1)(2(n-1)+1)}{6} - n\frac{(n-1)^2}{4}\right) \\
&= nb^2M^2\left(\frac{2n-1}{6} - \frac{n-1}{4}\right) \\
&= nb^2M^2\frac{4n-2-3n+3}{12} \\
&= n(n+1)\frac{b^2M^2}{12}
\end{aligned}$$

ovvero ha lo stesso valore per tutti i campioni $\mathbf{s}_1, \ldots, \mathbf{s}_M$. Lo stimatore (10.17) è quindi eguale a

$$\widehat{V} = \left(1 - \frac{n}{N}\right)(n+1)\frac{b^2M^2}{12} \tag{10.18}$$

qualunque sia il campione $\mathbf{s}_1, \ldots, \mathbf{s}_M$. Anche il suo valore atteso, ovviamente, è eguale alla (10.18). □

Esercizi

10.1. Data una popolazione di N unità $u_1, \ldots, u_N$, si supponga di scegliere in modo "casuale" una permutazione di $(1, \ldots, N)$, in modo tale che ognuna delle $N!$ permutazioni ha la stessa probabilità di essere selezionata. Se $(i_1, \ldots, i_N)$ è la permutazione scelta, all'unità u_{i_1} viene data etichetta 1,

all'unità u_{i_2} etichetta 2, e così via. Una volta assegnate le etichette, si supponga di scegliere dalla popolazione un campione sistematico di n unità (con $M = N/n$ intero). Provare che il campione contiene n qualunque unità della popolazione con probabilità $1/\binom{N}{n}$.

Suggerimento. Alle etichette $\{i, i+M, \ldots, i+(n-1)M\}$ viene assegnato un sottoinsieme di n qualsiasi unità della popolazione, e tutti i sottoinsiemi hanno la stessa probabilità di essere assegnati alle etichette in questione.

10.2. Provare che $\sum_{j=1}^{n} j = n(n+1)/2$.

Suggerimento. $1+2+3+\cdots+n = (1+n)+(2+(n-1))+(3+(n-2))+\cdots+(n+1) = (n+1)+(n+1)+(n+1)+\cdots+(n+1)$.

10.3. Provare che $\sum_{j=1}^{n} j^2 = n(n+1)(2n+1)/6$.

Suggerimento. $\sum_{j=1}^{n-1} j^3 = \sum_{j=1}^{n}(j-1)^3 = \sum_{j=1}^{n}(j^3 - 3j^2 + 3j - 1)$, da cui $n^3 = 3\sum_{j=1}^{n} j^2 - 3n(n+1)/2 + n$.

10.4. Per il generico campione $\mathbf{s}_i$ (10.13) ($i = 1, \ldots, M$), si definisca lo stimatore di $V(\overline{y}_{\mathbf{s}})$:

$$\widehat{V}_d = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{2(n-1)} \sum_{j=1}^{n-1} (y_{i+jM} - y_{i+(j-1)M})^2.$$

Provare le seguenti affermazioni.

a. Se l'assegnazione delle etichette alle unità è casuale, si ha

$$E[\widehat{V}_d] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2.$$

Suggerimento. Se l'assegnazione delle etichette alle unità è casuale, y_{i+jM}, $y_{i+(j-1)M}$ può essere una qualunque coppia di unità distinte della popolazione; ognuna di tali coppie ha probabilità $1/(N(N-1))$, per cui è $E[(y_{i+jM} - y_{i+(j-1)M})^2] = \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - y_j)^2/(N(N-1)) = 2S_y^2$.

b. Sotto l'ipotesi di trend lineare (10.9) si ha

$$E[\widehat{V}_d] = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{b^2 M^2}{2}.$$

Suggerimento. $\sum_{j=1}^{n-1}(y_{i+jM} - y_{i+(j-1)M})^2/(n-1) = b^2M^2$ qualunque sia il campione $\mathbf{s}_i$, $i = 1, \ldots, M$.

10.5. Una popolazione è formata da $N = 16$ unità, con etichette $i = 1, \ldots, 16$, e con valori y del tipo $y_i = i$, $i = 1, \ldots, 16$. Si considerino le seguenti due opzioni per la selezione di un campione di $n = 4$ unità.

– *Campionamento sistematico.* Si suddivide la popolazione negli $M = 4$ grappoli $\{1, 5, 9, 13\}$, $\{2, 6, 10, 14\}$, $\{3, 7, 11, 15\}$, $\{4, 8, 12, 16\}$, e si seleziona casualmente uno di essi. Si stima la media della popolazione con la media campionaria.

– *Campionamento stratificato.* Si suddivide la popolazione negli $M = 4$ strati $\{1, 5, 9, 13\}$, $\{2, 6, 10, 14\}$, $\{3, 7, 11, 15\}$, $\{4, 8, 12, 16\}$, e si seleziona casualmente una unità per ciascuno strato. Si stima la media della popolazione con la media campionaria.

Quale delle due opzioni è preferibile?

10.6. Per diventare ricco Pinocchio deve seppellire i suoi 5 denari nel campo dei miracoli, che ha superficie $100\,m^2$. Divide quindi il campo in parcelle quadrate di lato $1\,m$, e le numera secondo lo schema di seguito riportato.

1	2	3	4	5	6	7	8	9	10			
11	12	13	14	15	16	17	18	19	20			
21	22	23	24	25	26	27	28	29	30			
	31	32	33	34	35	36	37	38	39	40	41	42
	43	44	45	46	47	48	49	50	51	52	53	54
	55	56	57	58	59	60	61	62	63	64	65	66
	67	68	69	70	71	72	73	74	75	76		
			77	78	79	80	81	82	83	84		
			85	86	87	88	89	90	91	92		
			93	94	95	96	97	98	99	100		

Seppellisce quindi una moneta in ciascuno dei quadrati 5, 25, 45, 65, 85. Il gatto e la volpe, che non sanno né quanti denari abbia seppellito Pinocchio, né che schema abbia seguito, sono incerti sul da farsi. Il tempo a loro disposizione permette di scavare non più di 10 parcelle di terreno. Il gatto propone di selezionare le 10 parcelle da scavare mediante disegno sistematico. La volpe propone invece di selezionare le 10 parcelle da scavare mediante disegno ssr. Quale delle due strategie è preferibile?

11

Disegno campionario a due stadi semplici

11.1 Aspetti di base e notazione

In molti casi di interesse la popolazione si può pensare strutturata in grappoli, i quali potrebbero essere formati da parecchie unità elementari. Se si effettua un campionamento a grappolo, basta campionare pochi grappoli per avere un numero anche molto elevato di unità elementari. Ad esempio, si supponga di voler stimare il tasso di disoccupazione della popolazione di una città. Le unità elementari di rilevazione sono i singoli individui (diciamo di età 14–65 anni). In maniera molto semplice, si può pensare ad essi come raggruppati in grappoli, ciascuno dei quali è costituito da tutti gli individui che vivono in una stessa strada o piazza. Selezionare un campione di strade/piazze e osservare tutti gli individui che vivono negli edifici corrispondenti è un'operazione poco conveniente in quanto, a meno che la numerosità campionaria non sia molto grande, basteranno poche strade/piazze per raggiungere parecchie centinaia o migliaia di individui.

Ciò, come facilmente si intuisce, può avere conseguenze anche molto negative sull'efficienza di stima della media della popolazione. Un possibile rimedio consiste nel non osservare tutte le unità elementari dei grappoli selezionati, ma solo una parte di esse. È questa l'idea-guida del *disegno campionario a due stadi*. In particolare, in questo capitolo ci occuperemo esclusivamente del disegno a due stadi *semplici*. Esso parte, in buona sostanza, dalle stesse premesse del disegno a grappolo. Anche in questo caso il punto di partenza è una popolazione suddivisa in grappoli di unità elementari. Tuttavia, rispetto al disegno a grappolo, si ha un passo di campionamento aggiuntivo, nel senso che:

- dalla popolazione si seleziona, mediante disegno ssr, un *campione di grappoli*;
- da ciascuno dei grappoli selezionati si seleziona, sempre mediante disegno ssr, un campione di unità elementari.

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_11,

Chiaramente, il primo passo è identico a quello del disegno a grappolo. Tuttavia, *non si osservano tutte le unità elementari dei grappoli scelti, ma solo un loro campione*. Tale campione è ottenuto selezionando da ciascuno dei grappoli scelti un campione ssr di unità elementari.

La simbologia utilizzata è la stessa del Capitolo 9. Supporremo cioè che nella popolazione vi siano M grappoli, formati rispettivamente da N_1, N_2, ..., N_M unità elementari. Ciascuna unità elementare è individuata da una doppia etichetta (g, i), in cui $g\ (= 1, \ldots, M)$ è il grappolo a cui l'unità appartiene, e $i\ (= 1, \ldots, N_g)$ indica l'unità all'interno del grappolo di appartenenza. Si indicherà poi con $w_g = N_g/N$ il peso del grappolo g-mo ($g = 1, \ldots, M$).

Se y_{gi} è la modalità dell'unità $i\ (= 1, \ldots, N_g)$ del grappolo $g\ (= 1, \ldots, M)$, sempre seguendo la notazione in precedenza introdotta, si indicheranno con

$$\mu_{yg} = \frac{1}{N_g}\sum_{i=1}^{N_g} y_{gi}, \quad S^2_{yg} = \frac{1}{N_g - 1}\sum_{i=1}^{N_g}(y_{gi} - \mu_{yg})^2\,;\ g = 1, \ldots, M$$

rispettivamente la media e la varianza corretta del carattere di interesse $\mathcal{Y}$ nel grappolo g-mo. Per ogni grappolo g poniamo infine

$$z_g = M\, w_g\, \mu_{yg} = \frac{M}{N}\sum_{i=1}^{N_g} y_{gi}\,;\ g = 1, \ldots, M \tag{11.1}$$

così che, sempre similmente a quanto visto nei capitoli precedenti, la media della popolazione può essere espressa come:

$$\mu_y = \frac{1}{M}\sum_{g=1}^{M} z_g. \tag{11.2}$$

L'idea di base del disegno campionario a due stadi semplici è elementare.

- *I stadio.* Si seleziona, mediante campionamento ssr, un campione $\mathbf{g}_m$ di m degli M grappoli totali;
- *II stadio.* Da ciascun grappolo $g \in \mathbf{g}_m$ scelto al primo stadio si seleziona, mediante disegno ssr, un campione $\mathbf{s}_g$ di n_g unità elementari.

Fissato il campione $\mathbf{g}_m$ dei grappoli di primo stadio, la selezione di unità elementari avviene in modo indipendente nei diversi grappoli, così che gli m campioni $\mathbf{s}_g$, $g \in \mathbf{g}_m$, sono *indipendenti dato* $\mathbf{g}_m$.

Formalmente, ciascun campione di unità elementari si può scrivere come:

$$\mathbf{s} = \{\mathbf{s}_g;\ g \in \mathbf{g}_m\} = \{(g, i) \in \mathbf{s}_g;\ g \in \mathbf{g}_m\}.$$

Lo spazio dei campioni è pertanto l'insieme

$$\bigcup_{\mathbf{g}_m \in \mathcal{C}_{M,m}} \left\{ \bigtimes_{g \in \mathbf{g}_m} \mathcal{C}_{N_g, n_g} \right\}$$

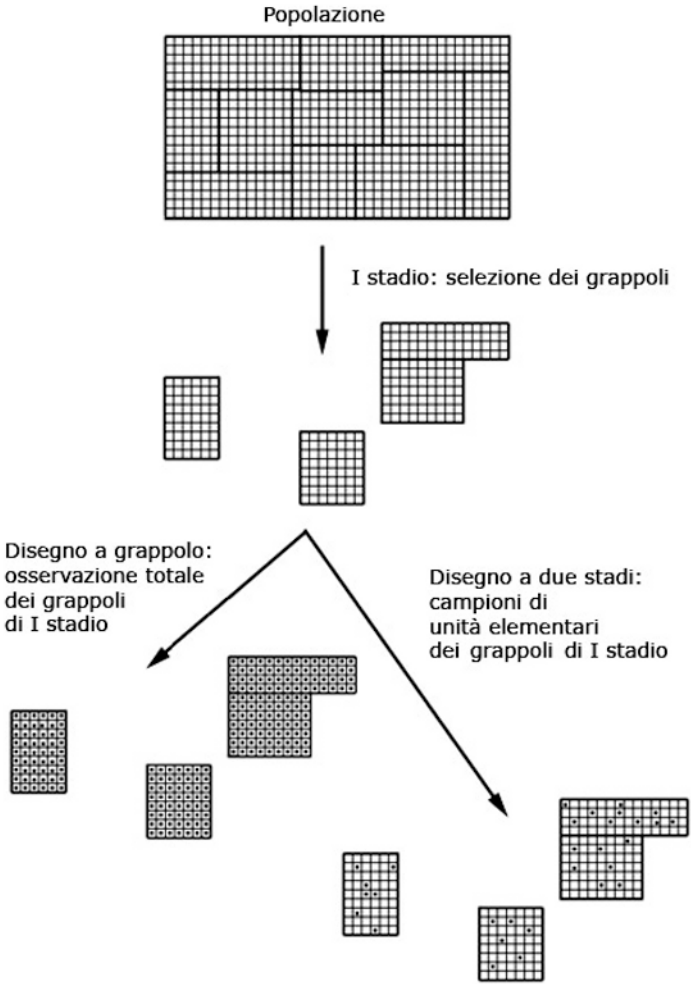


Fig. 11.1 Differenza tra disegno campionario a grappolo e a due stadi

e ciascun campione ha probabilità

$$P\left(\{\mathbf{s}_g;\ g \in \mathbf{g}_m\}\right) = P(\mathbf{g}_m)\, P(\{\mathbf{s}_g;\ g \in \mathbf{g}_m\}\,|\mathbf{g}_m) = \frac{1}{\binom{M}{m}} \prod_{g \in \mathbf{g}_m} \frac{1}{\binom{N_g}{n_g}}$$

dove

$$P(\mathbf{g}_m) = \frac{1}{\binom{M}{m}}$$

è la probabilità di selezionare al primo stadio il campione $\mathbf{g}_m$ di grappoli, mentre la

$$P(\{\mathbf{s}_g;\ g \in \mathbf{g}_m\}\,|\mathbf{g}_m) = \prod_{g \in \mathbf{g}_m} \frac{1}{\binom{N_g}{n_g}}$$

è la probabilità di selezionare al secondo stadio il campione $\{\mathbf{s}_g;\ g \in \mathbf{g}_m\}$ di unità elementari, condizionata all'aver scelto al primo stadio il campione $\mathbf{g}_m$ di grappoli.

Il disegno campionario a due stadi semplici comprende, come casi speciali, sia il disegno a grappolo che quello stratificato. In dettaglio, il disegno a grappolo si ottiene ponendo nel secondo stadio $n_1 = N_1, \ldots, n_M = N_M$, ossia selezionando tutte le unità elementari dei grappoli selezionati al primo stadio. Il disegno stratificato, invece, si ottiene considerando i grappoli come strati, e ponendo $m = M$ al primo stadio, ovvero selezionando tutti i grappoli da cui è formata la popolazione.

11.2 Considerazioni sul numero totale di unità elementari

In generale, se n_g è il numero di unità elementari selezionato dal grappolo g (scelto al primo stadio), il numero totale di unità elementari scelte con un disegno a due stadi (numerosità campionaria, per brevità) è pari a:

$$n_{tot} = \sum_{g \in \mathbf{g}_m} n_g. \tag{11.3}$$

Se le n_g sono date *a priori*, prima che venga selezionato il campione $\mathbf{g}_m$ dei grappoli di primo stadio, la (11.3) non è costante, a meno che le n_g non siano tutte uguali, ossia a meno che da ogni grappolo di primo stadio non si selezioni lo stesso numero di unità elementari. In ogni altro caso, n_{tot} dipende da quali grappoli sono selezionati al primo stadio, ossia da quali grappoli è composto $\mathbf{g}_m$. Posto $a_g = m\, n_g$, $g = 1, \ldots, m$, si vede facilmente che

$$n_{tot} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} m\, n_g = \frac{1}{m} \sum_{g \in \mathbf{g}_m} a_g$$

ovvero n_{tot} è la *media campionaria* dei numeri $a_1, \ldots, a_M$. Tenendo anche conto che il campione di grappoli $\mathbf{g}_m$ è selezionato con un disegno ssr, si ha quindi

$$E\,[n_{tot}] = \frac{1}{M} \sum_{g=1}^{M} a_g = \frac{m}{M}(n_1 + n_2 + \cdots + n_M). \tag{11.4}$$

Poiché il numero totale di unità campionarie selezionate, n_{tot} non è costante, non è possibile fissare *a priori* il numero totale n di unità da selezionare.

L'unica eccezione è quella in cui da ciascun grappolo si selezionano n/m unità elementari, così che $n_1 = \cdots = n_M = n/m$, e $n_{tot} = n$ qualunque sia il campione $\mathbf{g}_m$ di grappoli selezionati al primo stadio.

Esempio 11.1. Si supponga che la popolazione sia formata da $M = 4$ grappoli, rispettivamente di $N_1 = 100$, $N_2 = 300$, $N_3 = 100$; $N_4 = 200$ unità elementari. Si supponga inoltre che il disegno di campionamento sia a due stadi, e tale che:

- al primo stadio si selezionano $m = 2$ grappoli;
- dal grappolo 1 vengano selezionate $n_1 = 10$ unità, dal grappolo 2 $n_2 = 30$ unità, dal grappolo 3 $n_3 = 10$ unità, e dal grappolo 4 $n_4 = 20$ unità.

Nella Tabella 11.1 sono riportate le numerosità campionarie per i diversi campioni di grappoli selezionabili al primo stadio.

Tabella 11.1 Numerosità campionarie per disegno a due stadi

Campione $\mathbf{g}_m$ *di I stadio*	*Numerosità campionaria* n_{tot}
{1, 2}	$10 + 30 = 40$
{1, 3}	$10 + 10 = 20$
{1, 4}	$10 + 20 = 30$
{2, 3}	$30 + 10 = 40$
{2, 4}	$30 + 20 = 50$
{3, 4}	$10 + 20 = 30$

Il numero *medio* di unità campionarie elementari selezionate è pari a $2 \times (10 + 30 + 10 + 20)/4 = 35$. □

Se si vuole prefissare la numerosità campionaria n_{tot} pari a n, e non si vuole selezionare da ogni grappolo lo stesso numero di unità elementari, si deve agire in maniera completamente differente: il numero di unità elementari da selezionare da ciascun grappolo deve dipendere dal campione $\mathbf{g}_m$ dei grappoli di primo stadio. Siano $p_1, \ldots, p_M$ M numeri positivi (arbitrari, per il momento), tali che $p_1 + \cdots + p_M = 1$. Se si vuole che $n_{tot} = n$ qualunque sia il campione $\mathbf{g}_m$ dei grappoli scelti al primo stadio, allora da ciascun grappolo $g \in \mathbf{g}_m$ si deve scegliere un numero di unità elementari pari a:

$$n_g = n \frac{p_g}{\sum_{h \in \mathbf{g}_m} p_h}, \quad g \in \mathbf{g}_m. \tag{11.5}$$

Si osservi che quanto più grande è p_g, tanto più grande è n_g.

Per quanto riguarda i termini $p_1, \ldots, p_M$, una scelta abbastanza naturale (anche se non l'unica) consiste nel porre $p_g = w_g$, $g = 1, \ldots, M$. Inoltre, se $p_1 = \cdots = p_M$ $(= 1/M)$, allora da ogni grappolo si estrae lo stesso numero di unità, pari a n/m.

Esempio 11.2. Consideriamo la popolazione dell'Es. 11.1, formata da $M = 4$ grappoli, rispettivamente di $N_1 = 100$, $N_2 = 300$, $N_3 = 100$; $N_4 = 200$ unità elementari. Supponiamo anche che il disegno campionario sia a due stadi, e tale che al primo stadio si selezionano $m = 2$ grappoli. Supponiamo inoltre che $p_1 = 0.2$, $p_2 = 0.3$, $p_3 = 0.2$, $p_4 = 0.3$, e che si voglia una numerosità campionaria $n = 40$.

Nella Tabella 11.2 sono riportate le numerosità campionarie per i diversi campioni di grappoli selezionabili al primo stadio.

Tabella 11.2 Numerosità campionarie per disegno a due stadi ($n = 40$)

Campione $\mathbf{g}_m$ *di I stadio*	*Numerosità campionarie di II stadio*	
$\{1, 2\}$	$n_1 = 40 \times \frac{0.2}{0.2+0.3} = 16$	$n_2 = 40 \times \frac{0.3}{0.2+0.3} = 24$
$\{1, 3\}$	$n_1 = 40 \times \frac{0.2}{0.2+0.2} = 20$	$n_3 = 40 \times \frac{0.2}{0.2+0.2} = 20$
$\{1, 4\}$	$n_1 = 40 \times \frac{0.2}{0.2+0.3} = 16$	$n_4 = 40 \times \frac{0.3}{0.2+0.3} = 24$
$\{2, 3\}$	$n_2 = 40 \times \frac{0.3}{0.3+0.2} = 24$	$n_3 = 40 \times \frac{0.2}{0.3+0.2} = 16$
$\{2, 4\}$	$n_2 = 40 \times \frac{0.3}{0.3+0.3} = 20$	$n_4 = 40 \times \frac{0.3}{0.3+0.3} = 20$
$\{3, 4\}$	$n_3 = 40 \times \frac{0.2}{0.2+0.3} = 16$	$n_4 = 40 \times \frac{0.3}{0.2+0.3} = 24$

□

Nel seguito ci si concentrerà prevalentemente sul caso in cui $n_1, \ldots, n_M$ sono fissati *a priori*, più semplice sul piano degli sviluppi formali.

11.3 Stima della media della popolazione

Sulla base della (11.2) non è difficile costruire uno stimatore della media della popolazione. Nel caso del campionamento a grappolo l'osservazione di *tutte* le unità elementari dei grappoli selezionati permette di calcolare le medie μ_{yg} di tali grappoli, e quindi le quantità $z_g = M\, w_g\, \mu_{yg}$. La media campionaria di queste ultime porta ad uno stimatore "naturale" della media μ_y della popolazione.

Nel caso di disegno a due stadi questo non è possibile, in quanto essendo osservati solo campioni di unità dei grappoli di primo stadio, non è possibile calcolare le medie di tali grappoli. Un modo intuitivo per ovviare a tale inconveniente è quello di *stimare* le medie dei grappoli campionati al primo stadio con le corrispondenti *medie campionarie*. In simboli, se $\mathbf{g}_m$ è il campione di grappoli selezionati al primo stadio, e se da ogni grappolo $g \in \mathbf{g}_m$ si estrae un campione $\mathbf{s}_g$ di unità elementari, indichiamo con

$$\overline{y}_g = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi}, \quad g \in \mathbf{g}_m$$

la media campionaria del grappolo g. L'idea di base per stimare la media della popolazione, come già detto, consiste nello stimare le medie dei grappoli di

primo stadio con le corrispondenti medie campionarie, e quindi nel costruire uno stimatore simile a quello già visto nel caso di campionamento a grappolo, ma contenente appunto le medie campionarie in questione. Si ottiene in questo modo lo stimatore della media della popolazione:

$$\widehat{\mu}_{2st} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \overline{y}_g. \tag{11.6}$$

Le proprietà dello stimatore (11.6) sono delineate nelle successive proposizioni. Lo stimatore (11.6) è corretto sia quando le numerosità n_g di secondo stadio sono fissate *a priori* (e quindi n_{tot} è in generale variabile), sia quando sono scelte in base alla (11.5) (e quindi $n_{tot} = n$ fissato). Nel seguito, indicheremo con

$$\widehat{\mu}_{gr} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} z_g = \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \mu_{yg} \tag{11.7}$$

lo stimatore della media della popolazione basato su un disegno a grappolo (in cui, quindi, nel secondo stadio si osservano tutte le unità elementari dei grappoli scelti al primo stadio). Per semplicità indicheremo con il suffisso I il primo stadio di campionamento, e con II il secondo stadio di campionamento.

Proposizione 11.1. *Se il disegno campionario è a due stadi semplici, lo stimatore $\widehat{\mu}_{2st}$ possiede le seguenti due proprietà:*

- *il suo valore atteso (rispetto al II stadio)* condizionato *al campione di grappoli scelti al primo stadio è pari allo stimatore a grappolo* (11.7)*:*

$$E_{II}\left[\widehat{\mu}_{2st} \,|\mathbf{g}_m\right] = \widehat{\mu}_{gr}\,; \tag{11.8}$$

- *il valore atteso non condizionato è uguale alla media della popolazione:*

$$E_{I,\,II}\left[\widehat{\mu}_{2st}\right] = \mu_y \tag{11.9}$$

ossia $\widehat{\mu}_{2st}$ è uno stimatore corretto della media della popolazione.

Dimostrazione. Per provare la (11.8) è sufficiente tenere conto che il condizionare rispetto al campione $\mathbf{g}_m$ di primo stadio permette di trattare i grappoli g in $\mathbf{g}_m$ come fissati. Pertanto, sono fissate anche le numerosità campionarie n_g, $g \in \mathbf{g}_m$ di secondo stadio. Poichè la selezione di unità elementari dai grappoli g di $\mathbf{g}_m$ avviene mediante campionamento ssr, si ha $E[\overline{y}_g \,|\mathbf{g}_m] = \mu_{yg}$ per tutti i grappoli $g \in \mathbf{g}_m$. Quindi, si conclude che:

$$\begin{aligned} E_{II}\left[\widehat{\mu}_{2st} \mid \mathbf{g}_m\right] &= E_{II}\left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \overline{y}_g \mid \mathbf{g}_m\right] \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g E_{II}[\overline{y}_g \,|\mathbf{g}_m] \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \mu_{yg} \end{aligned}$$

ovvero la (11.8).

Per quanto riguarda la (11.9), è sufficiente usare la correttezza dello stimatore (11.7) (rispetto al primo stadio di campionamento) e una ben nota proprietà della media condizionata (*la media della media condizionata è uguale alla media non condizionata*):

$$\begin{aligned} E_{I,\,II}\left[\widehat{\mu}_{2st}\right] &= E_I\left[E_{II}\left(\widehat{\mu}_{2st}\,|\mathbf{g}_m\right)\right] = E_I\left[\widehat{\mu}_{gr}\right] \\ &= \mu_y. \end{aligned}$$

□

Il calcolo della varianza dello stimatore $\widehat{\mu}_{2st}$ è più complicato, e dipende dal criterio con cui sono determinate le numerosità campionarie n_g di secondo stadio. In questa sezione ci concentreremo esclusivamente sul caso in cui $n_1, \ldots, n_M$ sono fissate *a priori*. Con la stessa notazione del Capitolo 9, indichiamo con

$$S_b^2 = \frac{1}{M-1}\sum_{g=1}^{M}(z_g - \mu_y)^2 \tag{11.10}$$

la varianza corretta delle quantità z_g nella popolazione dei grappoli (con denominatore $M-1$ anziché M).

Proposizione 11.2. *Se il disegno campionario è a due stadi semplici con numerosità n_g di secondo stadio fissate* a priori*, si ha*

$$V_{I,\,II}\left(\widehat{\mu}_{2st}\right) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \frac{M}{m}\sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2. \tag{11.11}$$

Dimostrazione. Usando una ben nota proprietà di scomposizione della varianza (la varianza totale è pari alla somma della varianza della media condizionata e della media della varianza condizionata) si ha anzitutto:

$$V_{I,\,II}\left(\widehat{\mu}_{2st}\right) = V_I\left(E_{II}\left[\widehat{\mu}_{2st} \mid \mathbf{g}_m\right]\right) + E_I\left[V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right)\right]. \tag{11.12}$$

I due termini nel membro di destra della (11.12) possono essere calcolati separatamente. Per quanto riguarda il primo di essi, dalla (11.8) si ha:

$$V_I\left(E_{II}\left[\widehat{\mu}_{2st} \mid \mathbf{g}_m\right]\right) = V_I\left(\widehat{\mu}_{gr}\right) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2. \tag{11.13}$$

Per quanto riguarda invece il termine $E_I\left[V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right)\right]$, tenendo conto che *fissato* $\mathbf{g}_m$ *i campioni di secondo stadio* $\mathbf{s}_g$*,* $g \in \mathbf{g}_m$ *sono indipendenti e ssr*, si ha *in primis*

$$\begin{aligned} V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right) &= V_{II}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m} M w_g \overline{y}_g \,\middle|\, \mathbf{g}_m\right) \\ &= \frac{1}{m^2}\sum_{g\in\mathbf{g}_m} M^2 w_g^2 V_{II}\left(\overline{y}_g \mid \mathbf{g}_m\right) \\ &= \frac{1}{m^2}\sum_{g\in\mathbf{g}_m} M^2 w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2. \end{aligned}$$

Posto poi

$$d_g = \frac{M^2}{m} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{yg}^2, \quad g = 1, \ldots, M$$

e tenendo conto che il disegno di primo stadio è ssr sui grappoli, si ottiene

$$\begin{aligned} E_I \left[V_{II} \left(\widehat{\mu}_{2st} \mid \mathbf{g}_m \right) \right] &= E_I \left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} d_g \right] \\ &= \frac{1}{M} \sum_{g=1}^{M} d_g \\ &= \frac{1}{M} \sum_{g=1}^{M} \frac{M^2}{m} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{yg}^2 \\ &= \frac{M}{m} \sum_{g=1}^{M} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{yg}^2 \end{aligned} \tag{11.14}$$

Sommando infine le (11.13) e (11.14) si ottiene la varianza (11.11). □

La stima della varianza dello stimatore $\widehat{\mu}_{2st}$, sempre nel caso in cui le numerosità campionarie dei grappoli $n_1, \ldots, n_M$ siano fissate *a priori*, è abbastanza complessa. La costruzione di uno stimatore corretto di $V(\widehat{\mu}_{2st})$ è effettuata "per gradi" nella successiva Proposizione 11.3. Sia

$$\widehat{s}_{yg}^2 = \frac{1}{n_g - 1} \sum_{i \in \mathbf{s}_g} (y_{gi} - \overline{y}_g)^2, \quad g \in \mathbf{g}_m \tag{11.15}$$

la varianza campionaria corretta del grappolo g-mo, e poniamo:

$$\widehat{V}_1 = \frac{1}{m-1} \sum_{g \in \mathbf{g}_m} (M w_g \overline{y}_g - \widehat{\mu}_{2st})^2; \tag{11.16}$$

$$\widehat{V}_2 = \frac{1}{m} \sum_{g \in \mathbf{g}_m} w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g} \right) \widehat{s}_{yg}^2; \tag{11.17}$$

$$\widehat{V}_{2st} = \left(\frac{1}{m} - \frac{1}{M} \right) \widehat{V}_1 + M \widehat{V}_2. \tag{11.18}$$

Proposizione 11.3. *Se il disegno campionario è a due stadi semplici con numerosità n_g di secondo stadio fissate* a priori*, si ha*

$$E_{II} \left[\widehat{V}_1 \,\middle|\, \mathbf{g}_m \right] = \frac{1}{m-1} \sum_{g \in \mathbf{g}_m} (M w_g \mu_{yg} - \widehat{\mu}_{gr})^2 + m V_{II} \left(\widehat{\mu}_{2st} \,|\mathbf{g}_m \right); \tag{11.19}$$

$$E_{I,\,II} \left[\widehat{V}_1 \right] = S_b^2 + m E_I \left[V_{II} \left(\widehat{\mu}_{2st} \,|\, \mathbf{g}_m \right) \right]; \tag{11.20}$$

$$E_{I,\,II} \left[\widehat{V}_2 \right] = \frac{m}{M^2} E_I \left[V_{II} \left(\widehat{\mu}_{2st} \,|\, \mathbf{g}_m \right) \right]; \tag{11.21}$$

$$E_{I,\,II} \left[\widehat{V}_{2st} \right] = V(\widehat{\mu}_{2st}). \tag{11.22}$$

Dimostrazione. Per provare la (11.19), iniziamo con l'osservare che

$$\sum_{g\in\mathbf{g}_m}(Mw_g\overline{y}_g-\widehat{\mu}_{2st})^2=\sum_{g\in\mathbf{g}_m}(Mw_g\overline{y}_g-\widehat{\mu}_{gr})^2-m(\widehat{\mu}_{2st}-\widehat{\mu}_{gr})^2$$

da cui si ottiene, tenendo conto della (11.8),

$$\begin{aligned}
&E_{II}\left[\widehat{V}_1\,\middle|\,\mathbf{g}_m\right]\\
&=\frac{1}{m-1}E_{II}\left[\sum_{g\in\mathbf{g}_m}(Mw_g\overline{y}_g-\widehat{\mu}_{gr})^2\,\middle|\,\mathbf{g}_m\right]-\frac{m}{m-1}E_{II}\left[(\widehat{\mu}_{2st}-\widehat{\mu}_{gr})^2\,|\,\mathbf{g}_m\right]\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}E_{II}\left[(Mw_g\overline{y}_g-\widehat{\mu}_{gr})^2\,|\,\mathbf{g}_m\right]-\frac{m}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}\left\{V_{II}\left(Mw_g\overline{y}_g\,|\,\mathbf{g}_m\right)+(E_{II}[Mw_g\overline{y}_g\,|\mathbf{g}_m]-\widehat{\mu}_{gr})^2\right\}\\
&\quad-\frac{m}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}(Mw_g\mu_{yg}-\widehat{\mu}_{gr})^2+\frac{1}{m-1}V_{II}\left(\sum_{g\in\mathbf{g}_m}Mw_y\overline{y}_g\,\middle|\,\mathbf{g}_m\right)\\
&\quad-\frac{m}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}(Mw_g\mu_{yg}-\widehat{\mu}_{gr})^2+\frac{m^2}{m-1}V_{II}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m}Mw_y\overline{y}_g\,\middle|\,\mathbf{g}_m\right)\\
&\quad-\frac{m}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}(Mw_g\mu_{yg}-\widehat{\mu}_{gr})^2+\frac{m^2}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&\quad-\frac{m}{m-1}V_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)\\
&=\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}(Mw_g\mu_{yg}-\widehat{\mu}_{gr})^2+mV_{II}\left(\widehat{\mu}_{2st}\,|\,\mathbf{g}_m\right)
\end{aligned}$$

ovvero la (11.19). La (11.20) è una conseguenza immediata della (11.19) e del fatto che, come visto nel Capitolo 9,

$$E_I\left[\frac{1}{m-1}\sum_{g\in\mathbf{g}_m}(Mw_g\mu_{yg}-\widehat{\mu}_{gr})^2\right]=S_b^2.$$

La (11.21) si prova immediatamente osservando che $E_{II}[\widehat{s}^2_{yg}\,|\mathbf{g}_m]=S^2_{yg}$ e ripetendo gli stessi calcoli fatti in precedenza. Infine, la (11.22) è una conseguenza immediata delle (11.20), (11.21). □

Esempio 11.3. Come nell'Es. 9.1, consideriamo il *file* `fam2051.txt`, in cui sono riportati i dati relativi al comune di Statlandia. Le unità elementari sono famiglie, le quali sono raggruppate in edifici (grappoli), ciascuno formato da 8 famiglie. In totale vi sono $M = 128$ grappoli (edifici). I pesi dei grappoli sono tutti uguali, e pari a $1/128$: $w_g = 1/128$, $g = 1, \ldots, 128$. Le variabili nel *file* `fam2051.txt` sono 27; il significato di ciascuna di esse, e la corrispondente codifica, è riportato nel *file* `istruzioni_fam2051.txt`. Qui siamo interessati alla stima della media del reddito totale percepito dalle famiglie nell'anno 2050. Non essendo disponibile una lista delle famiglie, si effettua un campionamento a due stadi in cui:

- al primo stadio si selezionano $m = 10$ grappoli;
- al secondo stadio, da ciascuno dei grappoli campionati al primo stadio si selezionano 4 famiglie.

In Tabella 11.3 sono riportate le grandezze necessarie per costruire lo stimatore $\widehat{\mu}_{2st}$ e per stimare la sua varianza.

Come stima del reddito medio da lavoro della popolazione si ha:

$$\widehat{\mu}_{2st} = \frac{1}{10}\left(\overline{y}_3 + \overline{y}_9 + \cdots + \overline{y}_{112}\right) = 63497.$$

Per quanto riguarda invece la stima della varianza di $\widehat{\mu}_{2st}$, da

$$\widehat{V}_1 = \frac{1}{9}\sum (M w_g \overline{y}_g - \widehat{\mu}_{2st})^2 = 1743971458,$$
$$\widehat{V}_2 = \frac{1}{10}\sum \frac{1}{128^2}\left(\frac{1}{4} - \frac{1}{8}\right)\widehat{s}^2_{yg} = 434$$

si ottiene

$$\widehat{V}_{2st} = \left(\frac{1}{10} - \frac{1}{128}\right)\widehat{V}_1 + 128\widehat{V}_2 = 160827921. \qquad \square$$

Tabella 11.3 Costruzione degli stimatori $\widehat{\mu}_{2st}$ e $\widehat{V}_{2st}$

Grappoli	*Unità elementari*	$\overline{y}_g\ (= M w_g \overline{y}_g)$	$\widehat{s}^2_{yg}$	$(M w_g \overline{y}_g - \widehat{\mu}_{2st})^2$
3	1 2 7 8	148835	37958463	7282488906
9	1 3 5 8	108228	100958489	578667080
22	3 4 6 7	87553	100958489	578667080
35	1 6 7 8	69305	4624795	33729960
51	1 4 5 6	59883	14103536	13059189
62	2 3 5 6	47976	5628365	240901441
91	1 2 6 8	35814	35401826	1279010051
95	3 4 5 6	25868	21805223	1415941641
108	2 5 6 7	26814	145331844	1345642489
112	1 2 4 5	24697	197800671	1505440000

I risultati della proposizione precedente consentono anche di costruire stimatori corretti dei due termini

$$S_b^2 = \frac{1}{M-1}\sum_{g=1}^{M}(Mw_g\mu_{yg} - \mu_y)^2, \quad \sum_{g=1}^{M} w_g^2\left(\frac{1}{n_g} - \frac{1}{N_g}\right)S_{yg}^2.$$

In dettaglio, è facile verificare (Esercizio 11.1) che

$$E\left[\widehat{V}_1 - \frac{M}{m}\widehat{V}_2\right] = S_b^2; \tag{11.23}$$

$$E\left[M\widehat{V}_2\right] = \sum_{g=1}^{M} w_g^2\left(\frac{1}{n_g} - \frac{1}{N_g}\right)S_{yg}^2. \tag{11.24}$$

La costruzione di intervalli di confidenza per μ_y, infine, segue linee del tutto simili a quelle dei capitoli precedenti. Se il numero m di grappoli campionati al primo stadio è sufficientemente grande, e se il numero di unità elementari selezionate al secondo stadio da ciascuno dei grappoli di primo stadio è abbastanza grande, lo stimatore $\widehat{\mu}_{2st}$ ha distribuzione approssimata di tipo normale, con media μ_y e varianza $V(\widehat{\mu}_{2st})$. Con gli stessi ragionamenti dei capitoli precedenti, e sostituendo l'incognita $V(\widehat{\mu}_{2st})$ con la sua stima (11.18), si ha che la distribuzione di probabilità di

$$\frac{\widehat{\mu}_{2st} - \mu_y}{\widehat{V}_{2st}} \tag{11.25}$$

ha distribuzione approssimata di tipo normale standard. Detto pertanto, come al solito, z_α il quantile di ordine α della distribuzione normale standard, è immediato verificare che

$$\left[\widehat{\mu}_{2st} - z_{\alpha/2}\sqrt{\widehat{V}_{2st}},\ \widehat{\mu}_{2st} + z_{\alpha/2}\sqrt{\widehat{V}_{2st}}\right] \tag{11.26}$$

è un intervallo di confidenza per μ_y al livello approssimato $1-\alpha$.

Esempio 11.4. Consideriamo ancora l'Esempio 11.3. Visto l'esiguo numero di famiglie selezionate da ogni grappolo, l'approssimazione normale per la (11.25) non sarà probabilmente molto accurata. Ad ogni modo, a puro titolo di esempio numerico costruiamo l'intervallo (11.26) al livello di confidenza 0.95. Essendo $\sqrt{\widehat{V}_{2st}} = 12682$ e $z_{0.025} = 1.96$, si ha che l'intervallo

$$[63497 - 1.96\,12682,\ 63497 + 1.96\,12682] = [38641,\ 88353]$$

è un intervallo di confidenza approssimato per μ_y al livello 0.95. □

11.4 Caso speciale: grappoli della stessa numerosità

11.4.1 Aspetti di base

Esattamente come nel caso di disegno a grappolo, un caso speciale molto importante, e che merita una trattazione separata, è quello in cui gli M grappoli sono tutti formati dallo stesso numero L di unità:

$$N_1 = N_2 = \cdots = N_M = L.$$

Chiaramente si ha $N = M\,L$, e i pesi w_g sono tutti pari a $1/M$:

$$w_g = \frac{L}{N} = \frac{1}{M}, \quad g = 1, \ldots, M.$$

Le quantità z_g in (11.1) sono pertanto pari alle medie dei grappoli:

$$z_g = \mu_{yg}, \quad g = 1, \ldots, M.$$

In queste circostanze, è ragionevole prendere le numerosità campionarie dei grappoli tutte uguali, ponendo:

$$n_1 = n_2 = \cdots = n_M = l.$$

Il numero totale di unità elementari campionate, pertanto, è costante, ed eguale a:

$$n_{tot} = \sum_{g \in \mathbf{g}_m} l = ml = n. \tag{11.27}$$

Lo stimatore (11.6) si riduce alla media campionaria, in quanto:

$$\begin{aligned}
\widehat{\mu}_{2st} &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M \frac{1}{M} \overline{y}_g \\
&= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{1}{l} \sum_{i \in \mathbf{s}_g} y_{gi} \\
&= \frac{1}{ml} \sum_{g \in \mathbf{g}_m} \sum_{i \in \mathbf{s}_g} y_{gi} \\
&= \overline{y}_{\mathbf{s}}.
\end{aligned}$$

La varianza di $\widehat{\mu}_{2st}$ assume una forma molto semplificata. Usando infatti le stesse argomentazioni del Capitolo 9, valgono le due relazioni

$$S_b^2 = \frac{1}{M-1} \sum_{g=1}^{M} (\mu_{yg} - \mu_y)^2 \tag{11.28}$$

$$S_w^2 = \sum_{g=1}^{M} \frac{1}{M} S_{yg}^2 = \frac{1}{M(L-1)} \sum_{g=1}^{M} \sum_{i=1}^{L} (y_{gi} - \mu_{yg})^2 \tag{11.29}$$

dove la (11.28) è la *varianza tra i grappoli*, e la (11.29) è la *varianza nei grappoli*. La (11.11) si riduce quindi a:

$$V(\widehat{\mu}_{2st}) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \frac{M}{m} \sum_{g=1}^{M} \frac{1}{M^2} \left(\frac{1}{l} - \frac{1}{L}\right) S_{yg}^2$$
$$= \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \left(\frac{1}{ml} - \frac{1}{mL}\right) S_w^2. \tag{11.30}$$

Per quanto riguarda la stima della varianza di $\widehat{\mu}_{2st}$, valgono ovviamente i risultati già visti nella sezione precedente. Lo stimatore (11.18) assume ora una forma semplificata (Esercizio 11.2), ovvero:

$$\widehat{V}_{2st} = \left(\frac{1}{m} - \frac{1}{M}\right) \frac{1}{m-1} \sum_{g \in \mathbf{g}_m} (\overline{y}_g - \widehat{\mu}_{2st})^2 + \left(\frac{1}{ml} - \frac{1}{mL}\right) \frac{1}{M} \sum_{g \in \mathbf{g}_m} \widehat{s}_{yg}^2. \tag{11.31}$$

I risultati ottenuti fino ad ora permettono anche di ottenere stimatori corretti delle due varianze S_b^2 e S_w^2. In particolare, posto

$$\widehat{s}_b^2 = \frac{1}{m-1} \sum_{g \in \mathbf{g}_m} (\overline{y}_g - \widehat{\mu}_{2st})^2 - \left(\frac{1}{ml} - \frac{1}{mL}\right) \sum_{g \in \mathbf{g}_m} \widehat{s}_{yg}^2, \tag{11.32}$$

$$\widehat{s}_w^2 = \frac{1}{m} \sum_{g \in \mathbf{g}_m} \widehat{s}_{yg}^2 \tag{11.33}$$

è facile verificare (Esercizio 11.3) che valgono le relazioni

$$E\left[\widehat{s}_b^2\right] = S_b^2; \tag{11.34}$$

$$E\left[\widehat{s}_w^2\right] = S_w^2. \tag{11.35}$$

11.4.2 L'effetto del disegno

Una componente rilevante dell'effetto del disegno è dovuta alla presenza di grappoli di unità elementari. Per semplicità ci limiteremo al caso di grappoli tutti composti da L unità elementari. Il disegno considerato è esattamente quello visto in precedenza, ossia a due stadi con selezione di m grappoli al primo stadio, e di l unità elementari da ciascuno di essi. Come già rimarcato, lo stimatore $\widehat{\mu}_{2st}$ si riduce alla media campionaria: $\widehat{\mu}_{2st} = \overline{y}_{\mathbf{s}}$. La sua varianza, inoltre, è data dalla (11.30).

Assumendo che il fattore di correzione per popolazioni finite al primo stadio sia trascurabile, si può facilmente dimostrare (Esercizio 11.8) che l'effetto del disegno risulta pari a

$$Deff\,(2st,\ \overline{y}_{\mathbf{s}}) \approx 1 + (l-1)\rho_{ic} \tag{11.36}$$

dove ρ_{ic} è il coefficiente di correlazione intra-classi introdotto nel Capitolo 9. Se $l = L$ il disegno a due stadi si riduce a quello a grappolo, e la (11.36) di riduce all'effetto del disegno a grappolo.

Se i grappoli fossero formati casualmente, allora $\rho_{ic} \approx 0$ e il disegno a due stadi (a parità di numero totale di unità campionarie) sarebbe essenzialmente equivalente a quello semplice senza ripetizioni. Poiché in molti casi i grappoli non vengono formati da chi estrae il campione, ma sono piuttosto gruppi preesistenti di unità della popolazione, generalmente ρ_{ic} risulta positivo in quanto le unità all'interno di uno stesso grappolo tendono in genere ad essere più simili rispetto a unità di grappoli differenti. Ciò implica che il termine $Deff(2st,\ \overline{y}_{\mathbf{s}})$ risulta generalmente maggiore di 1, e quindi il campionamento a due stadi comporta una perdita di precisione rispetto ad un campionamento casuale semplice senza ripetizioni. Viceversa, un valore negativo del coefficiente di correlazione intra-classi comporta un effetto del disegno minore di uno, e quindi un incremento in precisione del campionamento a due stadi rispetto al campionamento casuale semplice.

La (11.36) mostra che per una data dimensione complessiva del campione ($n = lm$) l'effetto del disegno decresce al decrescere della dimensione dei campioni di secondo stadio. Se da una parte la riduzione della dimensione campionaria comporta una diminuzione dell'effetto del disegno, dall'altra occorrerà aumentare il numero di grappoli estratti, con un conseguente aumento dei costi di rilevazione dovuti alla dispersione territoriale delle unità del campione.

Esempio 11.5. Con riferimento alla (11.36), supponiamo che la popolazione sia suddivisa in 80 grappoli di dimensione 500. Supponiamo inoltre di estrarre un campione casuale semplice di $m = 10$ grappoli, e da ciascuno di essi un campione casuale semplice di $l = 100$ unità. Si hanno in totale $n = ml = 1000$ osservazioni campionarie. Sia inoltre $\rho_{ic} = 0.01$. L'effetto del disegno risulterà allora pari a 2. Ciò significa che il disegno semplice risulta circa due volte più preciso di quello a due stadi della stessa dimensione. Chiaramente, affinché la formula (11.36) sia utilizzabile, è necessario stimare preliminarmente la quantità ρ_{ic}. Un'alternativa pratica consiste nell'utilizzo di stime provenienti da precedenti indagini riguardanti la stessa variabile o variabili simili. L'effetto del disegno ci permette di stimare la dimensione che dovrebbe avere un campione casuale semplice per raggiungere la stessa precisione di quello a due stadi. Formalmente:

$$n_{eff}(2st,\ \overline{y}_{\mathbf{s}}) = \frac{1000}{2} = 500. \tag{11.37}$$

La (11.37) mostra che gli stessi risultati, in termini di precisione della stima, che abbiamo ottenuto con un campione a due stadi di 1000 unità si sarebbero potuti ottenere con un campione casuale semplice di sole 500 unità. □

Le indagini campionarie, soprattutto quelle svolte nell'ambito della statistica ufficiale, utilizzano generalmente disegni campionari complessi, caratterizzati da due o più stadi di campionamento, dalla stratificazione delle unità statistiche, dalla presenza di unità raggruppate in grappoli. Spesso, in aggiunta, la selezione delle unità di primo o dei successivi stadi è effettuata con

disegni di tipo non semplice, a probabilità variabile (essi verranno presentai nei Capitoli 12, 15). La complessità del disegno campionario provoca degli effetti sulla variabilità campionaria delle stime, che vengono valutati dall'*effetto del disegno*. Esso rappresenta quindi l'effetto cumulativo sulla precisione delle stime di aspetti del disegno, quali stratificazione, presenza di grappoli, utilizzo di probabilità variabili di selezione.

Se ad esempio il disegno è caratterizzato dalla presenza sia di strati sia di grappoli di unità, non possiamo affermare *a priori* se l'effetto del disegno sarà maggiore o minore di 1. Infatti, se da una parte la stratificazione tende ad aumentare la precisione delle stime, dall'altra la presenza di grappoli di unità tende a diminuirla.

11.5 Stima nel caso di numerosità totale costante*

Nel caso in cui le numerosità campionarie dei grappoli vengano scelte in base alla (11.5), la struttura della varianza dello stimatore $\widehat{\mu}_{2st}$ è più complicata. In dettaglio, vale la seguente proposizione.

Proposizione 11.4. *Se il disegno campionario è a due stadi semplici,* $\widehat{\mu}_{2st}$ *con numerosità* n_g *di secondo stadio scelte in base alla* (11.5)*, si ha*

$$V_I\left(E_{II}\left[\widehat{\mu}_{2st} \mid \mathbf{g}_m\right]\right) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2; \tag{11.38}$$

$$E_I\left[V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right)\right] = \frac{M}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\sum_{g=1}^{M} w_g^2 S_{yg}^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\sum_{g=1}^{M}\frac{w_g^2}{p_g} S_{yg}^2 - \frac{M}{mN}\sum_{g=1}^{M} w_g S_{yg}^2; \tag{11.39}$$

$$V_{I,\,II}\left(\widehat{\mu}_{2st}\right) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \frac{M^2}{n(M-1)}\left(\frac{1}{m} - \frac{1}{M}\right)\sum_{g=1}^{M} w_g^2 S_{yg}^2 - \frac{M}{mN}\sum_{g=1}^{M} w_g S_{yg}^2 + \frac{1}{n}\,\frac{M(m-1)}{m(M-1)}\sum_{g=1}^{M}\frac{w_g^2}{p_g} S_{yg}^2. \tag{11.40}$$

Dimostrazione. La varianza di $\widehat{\mu}_{2st}$ si può ancora scomporre in base alla (11.12), e il primo termine del membro di destra della (11.12) resta invariato:

$$V_I\left(E_{II}\left[\widehat{\mu}_{2st} \mid \mathbf{g}_m\right]\right) = V_I\left(\widehat{\mu}_{gr}\right) = \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2.$$

Per quanto riguarda la (11.39), si ha in primo luogo

$$\begin{aligned}
&V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right) \\
&= \frac{1}{m^2} \sum_{g \in \mathbf{g}_m} M^2 w_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_{yg}^2 \\
&= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{M^2}{m} w_g^2 \frac{S_{yg}^2}{n_g} - \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{M^2}{m} w_g^2 \frac{S_{yg}^2}{N_g} \\
&= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{M^2}{mn} \frac{w_g^2}{p_g} S_{yg}^2 \left(\sum_{g \in \mathbf{g}_m} p_g\right) - \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{M^2}{mN} w_g S_{yg}^2 \\
&= \frac{M^2}{n} \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{w_g^2}{p_g} S_{yg}^2\right) \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g\right) - \frac{M^2}{mN} \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} w_g S_{yg}^2\right)
\end{aligned}$$

da cui, posto

$$u_g = \frac{w_g^2}{p_g} S_{yg}^2, \quad g = 1, \ldots, M \tag{11.41}$$

$$t_g = w_g S_{yg}^2, \quad g = 1, \ldots, M \tag{11.42}$$

si vede che

$$V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right) = \frac{M^2}{n} \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} u_g\right) \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g\right) - \frac{M^2}{mN} \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} t_g\right).$$

Complessivamente, si ha quindi

$$\begin{aligned}
E_I\left[V_{II}\left(\widehat{\mu}_{2st} \mid \mathbf{g}_m\right)\right] = &\frac{M^2}{n} E_I \left[\left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} u_g\right) \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g\right)\right] \\
&- \frac{M^2}{mN} E_I \left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} t_g\right].
\end{aligned} \tag{11.43}$$

Tenendo infine conto che il disegno di primo stadio è ssr, che le quantità che appaiono nella (11.43) sono medie campionarie di grandezze dipendenti dai grappoli, e usando i risultati del Capitolo 3 (Sez. 3.7), si conclude che

$$\begin{aligned}
\frac{M^2}{mN} E_I \left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} t_g\right] &= \frac{M^2}{mN} \frac{1}{M} \sum_{g=1}^{M} t_g \\
&= \frac{M}{mN} \sum_{g=1}^{M} w_g S_{yg}^2
\end{aligned} \tag{11.44}$$

e

$$\begin{aligned}
&\frac{M^2}{n} E_I \left[\left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} u_g \right) \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g \right) \right] \\
&= \frac{M^2}{n} \left\{ C_I \left(\frac{1}{m} \sum_{g \in \mathbf{g}_m} u_g, \frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g \right) + E_I \left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} u_g \right] E_I \left[\frac{1}{m} \sum_{g \in \mathbf{g}_m} p_g \right] \right\} \\
&= \frac{M^2}{n} \left\{ \left(\frac{1}{m} - \frac{1}{M} \right) \frac{1}{M-1} \sum_{g=1}^{M} u_g \left(p_g - \frac{1}{M} \right) + \frac{1}{M^2} \sum_{g=1}^{M} u_g \right\} \\
&= \frac{M^2}{n(M-1)} \left(\frac{1}{m} - \frac{1}{M} \right) \sum_{g=1}^{M} w_g^2 S_{yg}^2 + \frac{M(m-1)}{m(M-1)} \sum_{g=1}^{M} \frac{w_g^2}{p_g} S_{yg}^2 \qquad (11.45)
\end{aligned}$$

da cui si ottiene subito la (11.39). La (11.11), infine, è un'immediata conseguenza di (11.38), (11.39). □

Per quanto riguarda la stima della varianza (11.11), valgono esattamente le considerazioni già fatte nella Sezione 11.3. Non è infatti difficile provare (Esercizio 11.4) che se si definiscono $\widehat{V}_1$, $\widehat{V}_2$, $\widehat{V}_{2st}$ rispettivamente come in (11.16), (11.17), (11.18), la Proposizione 11.3 continua a valere anche ora, pur di sostituire la (11.14) con la (11.39), e di usare per $V_{I\,II}(\widehat{\mu}_{2st})$ l'espressione (11.11). Del tutto simile a quella della Sezione 11.3, infine, è la costruzione di intervalli di confidenza per μ_y.

11.6 Grappoli di diversa numerosità e stimatore per quoziente

Il caso di grappoli di differente numerosià merita un'analisi un po' più approfondita di quella svolta fino ad ora, in quanto si presta a considerazioni del tutto simili a quelle svolte per il disegno a grappolo. Per semplicità ci limiteremo al caso in cui il numero di unità elementari da selezionare da ciascun grappolo sia fissato *a priori*. La varianza dello stimatore $\widehat{\mu}_{2st}$, come appare dalla (11.11), dipende da $S_b^2 = \sum_{g=1}^{M} (M w_g \mu_{yg} - \mu_y)^2/(M-1)$, ovvero dalla varianza (corretta) dei termini $M w_1 \mu_{y1}, \ldots, M w_m \mu_{yM}$. Detto

$$T_g = N_g \mu_{yg} = \sum_{i=1}^{N_g} y_{gi}$$

l'ammontare del carattere $\mathcal{Y}$ nel grappolo $g$mo, dalla relazione $M w_g \mu_{yg} = \frac{M}{N} T_g$ si vede subito che:

$$\begin{aligned}
S_b^2 &= Varianza\ di\ (M w_1 \mu_{y1}, \ldots, M w_M \mu_{yM}) \\
&\approx \left(\frac{M}{N} \right)^2 \times (Varianza\ di\ T_1, \ldots, T_M).
\end{aligned}$$

La varianza di $\widehat{\mu}_{2st}$ dipende quindi dalla varianza degli ammontari $T_1, \ldots, T_M$ dei grappoli. Quanto più grande è la loro variabilità, tanto più grande è il termine S_b^2, e quindi tanto più piccola è l'efficienza di $\widehat{\mu}_{2st}$. Si tratta di un discorso praticamente identico a quello già svolto per il campionamento a grappoli, e che quindi si presta a sviluppi e conclusioni simili.

Esattamente come nel caso del disegno a grappolo (e del relativo stimatore $\widehat{\mu}_{gr}$), lo stimatore $\widehat{\mu}_{2st}$ è altamente inefficiente quando i totali dei grappoli sono molto variabili come conseguenza di un'alta variabilità delle loro numerosità N_g, mentre le medie μ_{yg} dei grappoli sono abbastanza stabili. Questo equivale a dire che i totali dei grappoli possono essere considerati, con molta approssimazione, proporzionali alle relative numerosità: $T_g \approx cost\, N_g$, $g = 1, \ldots, M$, essendo *cost* un'opportuna costante di proporzionalità. L'esperienza empirica mostra che in molti casi di interesse questo è proprio ciò che accade: i totali $T_1, \ldots, T_M$ dei grappoli possiedono un'alta variabilità, ma soprattutto a causa della variabilità delle loro numerosità $N_1, \ldots, N_M$; le medie $\mu_{y1}, \ldots, \mu_{yM}$, per converso, sono relativamente "simili" l'una all'altra.

Per ovviare alla scarsa efficienza che lo stimatore $\widehat{\mu}_{2st}$ ha in casi come quello sopra descritto, è necessario mettere a punto un qualche stimatore alternativo, che possa fornire risultati migliori quando le numerosità dei grappoli sono molto variabili in presenza di medie dei grappoli stabili.

L'idea di base è di procedere esattamente come nel caso del disegno a grappolo (Sezione 9.4), costruendo uno stimatore di tipo quoziente in cui il ruolo di variabile ausiliaria è svolto dalla numerosità N_g dei grappoli. Formalmente, poniamo per ciascun grappolo g

$$x_g = N_g, \quad g = 1, \ldots, M.$$

La media di questa nuova variabile è pari a:

$$\mu_x = \frac{1}{M} \sum_{g=1}^{M} N_g = \frac{N}{M}\,.$$

In questo modo, si ha lo stimatore di "tipo quoziente" di μ_y:

$$\begin{aligned} \widehat{\mu}_{q2st} &= \frac{\frac{1}{m}\sum_{g\in\mathbf{g}_m} M w_g \overline{y}_g}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} x_g}\,\mu_x \\ &= \frac{\widehat{\mu}_{2st}}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} N_g}\,\frac{N}{M} \\ &= \frac{1}{M}\,\frac{\widehat{\mu}_{2st}}{\frac{1}{m}\sum_{g\in\mathbf{g}_m} w_g} \\ &= \frac{1}{M}\,\frac{\widehat{\mu}_{2st}}{\overline{w}_m} \end{aligned} \tag{11.46}$$

dove si è posto

$$\overline{w}_m = \frac{1}{m} \sum_{g\in\mathbf{g}_m} w_g = \textit{media campionaria dei pesi dei grappoli.}$$

Lo stimatore (11.46) ha un'efficienza (molto) superiore a quella dello stimatore $\widehat{\mu}_{2st}$ proprio nei casi in cui le medie dei grappoli sono "simili", mentre le numerosità N_g sono molto diverse tra loro. Ciò accade in quanto lo stimatore $\widehat{\mu}_{q2st}$ tende a controbilanciare la variabilità delle numerosità N_g dei grappoli con la media campionaria $\overline{w}_m$ al denominatore.

Le proprietà dello stimatore $\widehat{\mu}_{q2st}$ possono essere ottenute, in via approssimata, adattando la tecnica già vista nel caso del disegno a grappolo. Usando le stesse considerazioni già svolte per il disegno a grappolo, si può approssimare il termine $\widehat{\mu}_{q2st} - \mu_y$ nel seguente modo:

$$\begin{aligned}
\widehat{\mu}_{q2st} - \mu_y &= \frac{1}{M}\frac{\widehat{\mu}_{2st}}{\overline{w}_m} - \mu_y = \frac{\widehat{\mu}_{2st} - M\overline{w}_m\mu_y}{M\overline{w}_m} \\
&= \frac{1}{M\overline{w}_m}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g\overline{y}_g - M\mu_y\frac{1}{m}\sum_{g\in\mathbf{g}_m} w_g\right) \\
&= \frac{1}{M\overline{w}_m}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\overline{y}_g - \mu_y)\right) \\
&\approx \frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\overline{y}_g - \mu_y) \qquad (11.47)
\end{aligned}$$

con $\overline{w}_m \approx \sum_{g=1}^{M} w_g/M = 1/m$. Le proprietà (approssimate) di $\widehat{\mu}_{q2st}$ sono riassunte nella successiva proposizione.

Proposizione 11.5. *Se il disegno campionario è a due stadi semplici, con* $n_1, \ldots, n_M$ *fissati* a priori*, si ha:*

$$E_{II}\left[\widehat{\mu}_{q2st} \,|\, \mathbf{g}_m\right] = \widehat{\mu}_{qgr} \qquad (11.48)$$

$$E_{I,II}\left[\widehat{\mu}_{q2st}\right] \approx \mu_y \qquad (11.49)$$

$$\begin{aligned}
V_{I,II}(\widehat{\mu}_{q2st}) \approx &\left(\frac{1}{m} - \frac{1}{M}\right)\left\{\frac{M^2}{M-1}\sum_{g=1}^{M} w_g^2\left(\mu_{yg} - \mu_y\right)^2\right\} \\
&+\frac{M}{m}\sum_{g01}^{M} w_g^2 S_{yg}^2 \qquad (11.50)
\end{aligned}$$

dove $\widehat{\mu}_{qgr} = \frac{1}{M}\widehat{\mu}_{gr}/\overline{w}_m$ *è lo stimatore di tipo quoziente introdotto nella Sezione 9.4 nel caso di disegno a grappolo.*

Dimostrazione. La (11.48) si dimostra facilmente osservando che

$$\begin{aligned}
E_{II}\left[\widehat{\mu}_{q2st} \,|\, \mathbf{g}_m\right] &= \frac{1}{M}\,E_{II}\left[\left.\frac{\widehat{\mu}_{2st}}{\overline{w}_m}\right| \mathbf{g}_m\right] = \frac{1}{M}\frac{E_{II}\left[\widehat{\mu}_{2st} \,|\, \mathbf{g}_m\right]}{\overline{w}_m} \\
&= \frac{1}{M}\frac{\widehat{\mu}_{gr}}{\overline{w}_m} \\
&= \widehat{\mu}_{qgr}.
\end{aligned}$$

Per quanto riguarda il valore atteso approssimato di $\widehat{\mu}_{q2st}$, si ha poi

$$\begin{aligned}
E_{I,II}\left[\widehat{\mu}_{q2st}-\mu_y\right] &\approx E_{I,II}\left[\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\overline{y}_g-\mu_y)\right]\\
&= E_I\left[E_{II}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\overline{y}_g-\mu_y)\,\middle|\,\mathbf{g}_m\right)\right]\\
&= E_I\left[\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g\left(E_{II}\left[\overline{y}_g\,|\,\mathbf{g}_m\right]-\mu_y\right)\right]\\
&= E_I\left[\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\mu_{yg}-\mu_y)\right]\\
&= \frac{1}{M}\sum_{g=1}^{M} Mw_g(\mu_{yg}-\mu_y)\\
&= \sum_{g=1}^{M} w_g\mu_{yg}-\mu_y\\
&= 0
\end{aligned}$$

da cui la (11.49).

Per quanto riguarda infine la varianza approssimata dello stimatore $\widehat{\mu}_{q2st}$, inziamo con l'osservare che

$$\begin{aligned}
V_{I,II}(\widehat{\mu}_{q2st}) &= V_I\left(E_{II}\left[\widehat{\mu}_{q2st}\,|\,\mathbf{g}_m\right]\right)+E_I\left[V_{II}\left(\widehat{\mu}_{q2st}\,|\,\mathbf{g}_m\right)\right]\\
&= V_I\left(\widehat{\mu}_{qgr}\right)+E_I\left[V_{II}\left(\widehat{\mu}_{q2st}\,|\,\mathbf{g}_m\right)\right]. \qquad (11.51)
\end{aligned}$$

Il primo termine che compare al membro di destra della (11.51) è pari (vds. Sezione 9.4) a

$$V_I\left(\widehat{\mu}_{qgr}\right)\approx\left(\frac{1}{m}-\frac{1}{M}\right)\frac{M^2}{M-1}\sum_{g=1}^{M} w_g^2(\mu_{yg}-\mu_y)^2. \qquad (11.52)$$

Per il secondo termine, invece, osserviamo che

$$\begin{aligned}
V_{II}\left(\widehat{\mu}_{q2st}\,|\,\mathbf{g}_m\right) &\approx V_{II}\left(\frac{1}{m}\sum_{g\in\mathbf{g}_m} Mw_g(\overline{y}_g-\mu_y)\,\middle|\,\mathbf{g}_m\right)\\
&= \frac{1}{m^2}\sum_{g\in\mathbf{g}_m} M^2w_g^2V_{II}\left(\overline{y}_g\,|\,\mathbf{g}_m\right)\\
&= \frac{M^2}{m^2}\sum_{g\in\mathbf{g}_m} w_g^2\left(\frac{1}{n_g}-\frac{1}{N_g}\right)S_{yg}^2
\end{aligned}$$

da cui

$$E_I\left[V_{II}\left(\widehat{\mu}_{q2st}\,|\,\mathbf{g}_m\right)\right] \approx \frac{M^2}{m} E_I\left[\frac{1}{m}\sum_{g\in\mathbf{g}_m} w_g^2\left(\frac{1}{n_g}-\frac{1}{N_g}\right)S_{yg}^2\right]$$
$$= \frac{M}{m}\sum_{g=1}^{M} w_g^2\left(\frac{1}{n_g}-\frac{1}{N_g}\right)S_{yg}^2. \qquad (11.53)$$

Sommando infine (11.52) e (11.53), si ottiene la (11.50). □

Sulla base dei risultati precedenti, non è difficile costruire uno stimatore della varianza approssimata di $\widehat{\mu}_{q2st}$. Con la stessa notazione della Sezione 11.3, è infatti intuitivo fare riferimento a

$$\widehat{V}_{q2st} = \left(\frac{1}{m}-\frac{1}{M}\right)\frac{M^2}{m-1}\sum_{g\in\mathbf{g}_m} w_g^2\left(\overline{y}_g-\widehat{\mu}_{q2st}\right)^2 + \frac{M^2}{m^2}\sum_{g\in\mathbf{g}_m} w_g^2 s_{yg}^2. \quad (11.54)$$

Per quanto riguarda l'efficienza dello stimatore $\widehat{V}_{q2st}$, valgono considerazioni del tutto simili a quelle svolte nel caso del disegno campionario a grappolo. Dall'espressione (11.50) appare chiaro che la varianza di $\widehat{\mu}_{q2st}$ è tanto più piccola quanto più piccolo è il termine $\sum_{g=1}^{M} w_g^2\left(\mu_{yg}-\mu_y\right)^2/(M-1)$, ossia quanto più bassa è la variabilità delle medie μ_{yg} dei grappoli. In sostanza, lo stimatore per quoziente $\widehat{\mu}_{qgr}$ è tanto più efficiente quanto più le medie dei grappoli tendono ad essere "simili" tra loro.

11.7 Il problema della scelta del numero di grappoli e di unità elementari

Esattamente come nel caso del disegno a grappolo, il campionamento a due stadi è di frequente utilizzato in quanto le unità elementari di uno stesso grappolo sono fisicamente "vicine", e la loro osservazione non richiede di sostenere grossi costi di spostamento. Talvolta i grappoli sono suggeriti in modo naturale dall'oggetto della rilevazione. Altre volte, invece, vi è anche la possibilità di scegliere sia il numero M dei grappoli in cui viene suddivisa la popolazione, sia il numero di unità elementari da cui sono formati i grappoli. Su questo punto sono necessarie alcune considerazioni. Per semplicità ci si riferirà a caso in cui i grappoli sono tutti composti dallo stesso numero L di unità elementari, per cui se è dato il numero totale $N = ML$ di unità elementari della popolazione, scegliere L equivale a scegliere M. Ovviamente, la scelta dovrebbe essere effettuata in modo da rendere quanto più piccola possibile la varianza dello stimatore $\widehat{\mu}_{2st}$, che nel caso in esame è data dalla (11.30). Da un punto di vista intuitivo, quanto più grande è M tanto più piccolo S_b^2. Tuttavia, come visto nella Sezione 9.3, vale la relazione approssimata $S_y^2 \approx S_b^2 + S_w^2$, dalla quale discende che, fissato S_y^2, quanto più grande è M tanto più grande è S_w^2.

Vi è quindi un *trade-off* per la scelta di M, in quanto per rendere "piccola" la prima parte della (11.30) conviene scegliere M piccolo, mentre per rendere piccola la seconda parte della (11.30) conviene scegliere M grande. Per impostare formalmente il problema della scelta di M, m e l si potrebbe assumere un opportuno legame tra L $(= N/M)$ e S_w^2, e un vincolo sul costo totale di rilevazione. Poiché, tranne in casi eccezionali, i legami tra L e S_w^2 sono molto difficili da esplicitare, nel seguito ci si accontenterà di un obiettivo molto meno ambizioso (ma più realistico): dati i grappoli in cui è suddivisa la popolazione, si devono scegliere sia il numero m di grappoli da selezionare al primo stadio, e del numero di unità elementari da selezionare al secondo stadio. Purtroppo, a causa delle diverse grandezze da scegliere, questo problema *non* può essere impostato in maniera elementare prefissando l'errore massimo ammissibile di stima e la probabilità con cui si supera tale errore. Nel seguito si seguirà una strada completamente diversa, basata sulla minimizzazione della varianza di stima subordinatamente ad opportuni vincoli sul costo della rilevazione.

11.7.1 Grappoli tutti della stessa numerosità

Nel caso in cui la popolazione sia suddivisa in M grappoli tutti della stessa numerosità L, come già detto, è abbastanza naturale (benché non ottimale) selezionare da ciascuno degli m grappoli di primo stadio lo stesso numero l di unità elementari. Il problema consiste nello scegliere m e l. Come si vede facilmente a partire dalla (11.30), la varianza dello stimatore $\widehat{\mu}_{2st}$ si può esprimere in questo caso come:

$$V(\widehat{\mu}_{2st}) = \frac{1}{m}\left\{S_b^2 - \frac{S_w^2}{L} + \frac{S_w^2}{l}\right\} - \frac{S_b^2}{M}. \tag{11.55}$$

L'idea di base è quella di minimizzare la (11.55), subordinatamente ad un qualche vincolo sul costo totale di rilevazione. La più semplice funzione di costo è del tipo $C = c_1 m + c_2 ml$ dove c_1 è il costo di "contatto" per un grappolo, e c_2 è il costo di osservazione di un'unità elementare. Se C_0 è il *budget* a disposizione per la rilevazione, si ha quindi il vincolo:

$$c_1 m + c_2 ml = C_0. \tag{11.56}$$

L'idea di base per scegliere m e l è molto semplice: bisogna determinare i valori di m e l che minimizzano la (11.55), subordinatamente al vincolo (11.56). Poiché il termine S_b^2/M non dipende né da m, né da l, il problema di minimo vincolato da risolvere diviene:

$$\begin{cases} \textit{minimizzare}: \ \frac{1}{m}\left\{S_b^2 - \frac{S_w^2}{L} + \frac{S_w^2}{l}\right\} \\ \textit{con il vincolo}: \ c_1 m + c_2 ml = C_0 \end{cases}. \tag{11.57}$$

Usando il metodo dei moltiplicatori di Lagrange, e assumendo che $S_b^2 - S_w^2/L > 0$ è facile provare (Esercizio 11.5) che i valori ottimi di l e m sono eguali a:

$$l^* = \frac{1}{\sqrt{S_b^2/S_w^2 - 1/L}} \sqrt{\frac{c_1}{c_2}}, \quad m^* = \frac{C_0}{c_1 + l^* c_2}\,. \tag{11.58}$$

In generale, se da ciascun grappolo di secondo stadio si selezionano l unità elementari, il vincolo di *budget* in (11.56) implica che $m = C_0/(c_1 + lc_2)$. La varianza di $\widehat{\mu}_{2st}$ è pertanto pari a:

$$V(\widehat{\mu}_{2st}) = \frac{c_1 + lc_2}{C_0} \left(S_b^2 - \frac{S_w^2}{L} + \frac{S_w^2}{l} \right) - \frac{S_b^2}{M}. \tag{11.59}$$

In particolare, ponendo $l = l^*$, $m = m^*$ nella (11.59) si ottiene il valore ottimo della varianza dello stimatore $\widehat{\mu}_{2st}$.

Nella pratica i valori di S_b^2 e S_w^2 sono in genere incogniti. In tal caso si possono seguire diverse strade. In primo luogo, è da osservare che il valore ottimo di $V(\widehat{\mu}_{2st})$ è relativamente poco sensibile rispetto a l, almeno per un buon *range* di valori di l. In altre parole, se l non si discosta di molto dal suo valore ottimo l^*, si ha solo un piccolo aumento di $V(\widehat{\mu}_{2st})$. Pertanto, anche una conoscenza largamente approssimata dei valori del rapporto S_b^2/S_w^2 può condurre ad una buona scelta di l (e di conseguenza di m). Ciò è mostrato nel successivo esempio.

Esempio 11.6. Si considerino ancora i dati del *file* `fam2051.txt`, già visti negli Es. 9.1 e 11.1. La popolazione è composta da famiglie (unità elementari), raggruppate in edifici (grappoli) ciascuno composto da $L = 8$ famiglie; in totale, vi sono $M = 128$ grappoli. In particolare, si consideri la variabile "reddito totale nell'anno 2050". Le varianze nei grappoli e tra i grappoli sono rispettivamente pari a: $S_b^2 = 1343095048$, $S_w^2 = 117345278$, così che $S_b^2/S_w^2 = 11.45$.

Per quanto riguarda la funzione di costo, assumiamo che $c_1 = 50c_2$, $c_2 = 1$, e che il *budget* totale sia $C_0 = 2000$. Il valore ottimo della numerosità campionaria nei grappoli (ossia del numero di unità elementari da selezionare da ciascun grappolo di primo stadio) è pari a

$$l^* = \frac{1}{\sqrt{11.45} - 1/8} \sqrt{50} \approx 2$$

e di conseguenza il numero ottimo di grappoli-campione di primo stadio è

$$m^* = \frac{2000}{52} \approx 38.$$

La corrispondente varianza dello stimatore $\widehat{\mu}_{2st}$ è pari a:

$$\begin{aligned} V^*(\widehat{\mu}_{2st}) &= \frac{52}{2000} \left(1343095048 - \frac{117345278}{8} + \frac{117345278}{2} \right) - \frac{1343095048}{128} \\ &\approx 25571658. \end{aligned} \tag{11.60}$$

Tabella 11.4 Varianza di $\widehat{\mu}_{2st}$ per diversi valori di l

l	$V(\widehat{\mu}_{2st})$	$100 \times \left(\frac{V(\widehat{\mu}_{2st})}{V^*(\widehat{\mu}_{2st})} - 1\right)$
2	25571658	0.0
3	25746932	0.6
4	26166677	2.3
5	26684208	4.3
6	27250634	6.5
7	27844999	8.9
8	28456826	11.3

Nella successiva Tabella 11.4 sono riportati i valori di $V(\widehat{\mu}_{2st})$ per $l = 2 - 8(1)$ (e per i corrispondenti valori di $m = 2000/(50 + l)$, naturalmente), nonché le differenze relative rispetto all'ottimo (11.60).

Come appare evidente, la varianza di $\widehat{\mu}_{2st}$ è sempre molto vicina all'ottimo; in molti casi non supera l'ottimo del 5%, ed anche nel caso peggiore è dell'11% superiore all'ottimo. □

Una strada alternativa consiste nello stimare S_b^2 e S_w^2 mediante un campione pilota. L'idea-guida è elementare. Mediante un disegno a due stadi semplici, si seleziona al primo stadio un campione ssr di m_p grappoli, da ciscuno dei quali si seleziona un campione ssr di l_p unità elementari. Si stimano poi S_b^2 e S_w^2 rispettivamente con gli stimatori (11.34) e (11.35). Infine, si stimano i valori ottimi di l e m con:

$$\widehat{l}^* = \frac{1}{\sqrt{\widehat{s}_b^2/\widehat{s}_w^2 - 1/L}} \sqrt{\frac{c_1}{c_2}}, \quad m^* = \frac{C_0}{c_1 + \widehat{l}^* c_2}.$$

I valori di m_p, l_p dovrebbero essere "piccoli", ma tuttavia in grado di fornire stime sufficientemente accurate di S_b^2 e S_w^2. Ulteriori approfondimenti su questo ed altri aspetti sono nel volume di Cochran (1977), pp. 283–285.

In chiusura, un'ultima considerazione. L'idea di selezionare da ciascun grappolo di primo stadio lo *stesso* numero di unità elementari è basata sull'intuizione e sulla semplicità, ma non ha nessuna particolare giustificazione dal punto di vista dell'ottimalità della varianza dello stimatore $\widehat{\mu}_{2st}$. Un'idea alternativa è quella di scegliere il numero di unità da selezionare dai grappoli in modo da rendere minima la varianza dello stimatore $\widehat{\mu}_{2st}$. Quest'impostazione verrà usata nel prossimo paragrafo, con riferimento al caso in cui i grappoli non hanno tutti necessariamente la stessa numerosità. Ovviamente, come caso speciale, si può trattare anche il caso di grappoli tutti formati dallo stesso numero di unità elementari.

*11.7.2 Grappoli di diversa numerosità**

Nel caso di grappoli di differente numerosità, l'approccio è abbastanza simile a quello seguito in precedenza, anche se i risultati sono meno nitidi. Suppo-

niamo, al solito, che debbano essere stabiliti *a priori* $n_1, \ldots, n_M$, ovvero che debba essere stabilito *a priori* il numero di unità da selezionare da ciascun grappolo. Per semplicità di notazione poniamo

$$n_{sum} = n_1 + n_2 + \cdots + n_M$$
$$p_g = \frac{n_g}{n_{sum}}; \;\; g = 1, \ldots, M$$

così che

$$n_g = n_{sum}\, p_g; \;\; g = 1, \ldots, M$$

con la condizione

$$p_1 + p_2 + \cdots + p_M = 1. \tag{11.61}$$

Il problema di scegliere m, $n_1, \ldots, n_M$ equivale, ovviamente, al problema di scegliere m, n_{sum}, $p_1, \ldots, p_M$, con il vincolo (11.61).

La varianza dello stimatore $\widehat{\mu}_{2st}$ risulta pari a:

$$\begin{aligned} V(\widehat{\mu}_{2st}) &= \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \frac{M}{m} \sum_{g=1}^{M} w_g^2 \left(\frac{1}{p_g n_{sum}} - \frac{1}{N_g}\right) S_{yg}^2 \\ &= \left(\frac{1}{m} - \frac{1}{M}\right) S_b^2 + \frac{M}{m\, n_{sum}} \sum_{g=1}^{M} \frac{w_g^2}{p_g} S_{yg}^2 - \frac{M}{mN} \sum_{g=1}^{M} w_g S_{yg}^2 \\ &= \frac{1}{m} \left\{ S_b^2 - \frac{M}{N} \sum_{g=1}^{M} w_g S_{yg}^2 + \frac{M}{n_{sum}} \sum_{g=1}^{M} \frac{w_g^2}{p_p} S_{yg}^2 \right\} - \frac{1}{M} S_b^2. \end{aligned} \tag{11.62}$$

L'idea di base è molto semplice, e sostanzialmente simile a quella già usata nella sezione precedente: minimizzare la varianza (11.62), subordinatamente al vincolo (11.61) e ad un oppurtuno vincolo sul costo della rilevazione. Per quanto riguarda quest'ultimo, indichiamo con:

- c_1: costo di "contatto" per un grappolo;
- c_1: costo di "inserimento in lista" per un'unità elementare;
- c_2: costo di osservazione di un'unità elementare.

Nel caso di grappoli tutti della stessa numerosità il costo c_l di inserimento in lista è assente in quanto per i diversi grappoli si sfrutta sostazialmente la stessa lista. Il costo di rilevazione è ora del tipo $C = c_1 m + \sum_{g \in \mathbf{g}_m} n_g c_2 + \sum_{g \in \mathbf{g}_m} N_g c_l$, e chiaramente dipende dal campione $\mathbf{g}_m$ di primo stadio. Il *costo medio* di rilevazione è pari (Esercizio 11.6) a:

$$E[C] = m c_1 + \frac{m}{M} n_{sum} c_2 + \frac{m}{M} N c_l. \tag{11.63}$$

Se C_0 è il *budget* a disposizione per la rilevazione, si pone il seguente vincolo sul costo medio:

$$m c_1 + \frac{m}{M} n_{sum} c_2 + \frac{m}{M} N c_l = C_0. \tag{11.64}$$

Il problema da risolvere è quello della minimizzazione della varianza (11.62) subordinatamente ai vincoli (11.61), (11.64).

Scelta ottima di n_{sum}, m

Supponiamo per il momento che i valori di $p_1, \ldots, p_M$ nella (11.62) siano dati (la scelta dei valori ottimi di $p_1, \ldots, p_M$ verrà affrontata in seguito). Bisogna minimizzare tale varianza rispetto a n_{sum} e m, subordinatamente al vincolo (11.64). Il termine S_b^2/M si può trascurare, in quanto non dipende né da m, né da n_{sum}. Pertanto, il problema di minimo vincolato da risolvere è del tipo:

$$\begin{cases} \textit{minimizzare}: \ \frac{1}{m}\left\{S_b^2 - \frac{M}{N}\sum_{g=1}^{M} w_g S_{yg}^2 + \frac{M}{n_{sum}}\sum_{g=1}^{M}\frac{w_g^2}{p_g}S_{yg}^2\right\} \\ \textit{con il vincolo}: \ mc_1 + \frac{m}{M}n_{sum}c_2 + m\frac{N}{M}c_l = C_0 \end{cases}. \quad (11.65)$$

È facile provare (Esercizio 11.7) che i valori ottimi di n_{sum} e m sono rispettivamente eguali a:

$$n_{sum}^* = \sqrt{\frac{\sum_{g=1}^{M}\frac{w_g^2}{p_g}S_{yg}^2}{\frac{1}{M}S_b^2 - \frac{1}{N}\sum_{g=1}^{M} w_g S_{yg}^2}}\sqrt{\frac{Mc_1 + Nc_l}{c_2}}; \quad (11.66)$$

$$m^* = \frac{C_0}{c_1 + \frac{n_{sum}^*}{M}c_2 + \frac{N}{M}c_l}. \quad (11.67)$$

Scelta ottima di $p_1, \ldots, p_M$

Bisogna adesso determinare i valori ottimi di $p_1, \ldots, p_M$. Poiché l'unica parte della varianza (11.62) che dipende da $p_1, \ldots, p_M$ è il termine $\sum_{g=1}^{M}\frac{w_g^2}{p_p}S_{yg}^2$, per determinare i valori ottimi di $p_1, \ldots, p_M$ bisogna risolvere il seguente problema di minimo vincolato

$$\begin{cases} \textit{minimizzare}: \ \sum_{g=1}^{M}\frac{w_g^2}{p_g}S_{yg}^2 \\ \textit{con il vincolo}: \ p_1 + \cdots + p_M = 1 \end{cases}. \quad (11.68)$$

Si tratta di un problema quasi identico a quello dell'allocazione ottimale nel disegno stratificato. Usando le medesime tecniche, si ricava che i valori ottimi di $p_1, \ldots, p_M$ sono pari a:

$$p_g^* = \frac{w_g S_{yg}}{\sum_{h=1}^{M} w_h S_{yh}}; \ \ g = 1, \ldots, M. \quad (11.69)$$

Si vede facilmente che se si pone $p_1 = p_1^*, \ldots, p_M = p_M^*$, i valori ottimi di n_{sum} e m divengono pari a:

$$n_{sum}^{**} = \frac{\sum_{g=1}^{M} w_g S_{yg}}{\sqrt{\frac{S_b^2}{M} - \frac{1}{N}\sum_{g=1}^{M} w_g S_{yg}^2}}\sqrt{\frac{Mc_1 + Nc_l}{c_2}}; \quad (11.70)$$

$$m^{**} = \frac{C_0}{c_1 + \frac{n_{sum}^{**}}{M}c_2 + \frac{N}{M}c_l}. \quad (11.71)$$

Considerazioni sull'applicabilità dei risultati ottenuti

I valori ottimi (11.69), (11.70), (11.71) dipendono dalle quantità w_g (che potrebbero in alcuni casi essere incognite) e S^2_{yg} (che sono praticamente sempre incognite). Metodi per ovviare almeno in parte a questo inconveniente sono brevemente delineati nel seguito.

In primo luogo, se non si hanno informazioni sulle varianze S^2_{yg} dei grappoli è giocoforza rinunciare ai pesi ottimi (11.69), e ripiegare su soluzioni "ragionevoli". Se sono note le numerosità N_g dei grappoli, si potrebbe porre $p_g = w_g$, oppure $p_g = 1/M$ (quest'ultima scelta non necessita della conoscenza dei termini w_g, ed inoltre fornisce una numerosità campionaria costante). Se anche le numerosità N_g dei grappoli sono incognite *a priori*, e possono essere conosciute solo contattando i grappoli (ciò accade non di rado), la scelta $p_g = 1/M$ appare assai ragionevole.

Se si scelgono valori di p_g non ottimali, i valori di n_{sum}, m da utilizzare sono dati rispettivamente dalle (11.66), (11.67). Purtroppo, essi dipendono dalle incognite varianze dei grappoli, S^2_{yg}. Questo inconveniente può essere almeno in parte superato mediante la tecnica del campione pilota. L'idea-guida è simile a quella già descritta nel caso di grappoli tutti della stessa numerosità. Tramite un disegno a due stadi semplici, si seleziona al primo stadio un campione ssr di m_p grappoli, da ciascuno dei quali si seleziona un campione ssr di $n_{p,g}$ unità elementari (ad esempio, ma non necessariamente, da ciascun grappolo del campione pilota si potrebbe selezionare lo stesso numero di unità elementari). Indichiamo con $\overline{y}_{p\,g}$ $\widehat{s}^2_{p\,yg}$ rispettivamente le medie e le varianze campionarie dei grappoli del campione pilota. Si stima poi S^2_b con lo stimatore che appare al membro di sinistra della (11.23) (che indichiamo con $\widehat{s}^2_{p\,b}$), e si stimano $\sum_g w_g^2 S^2_{yg}/p_g$, $\sum_g w_g S^2_{yg}$ rispettivamente con

$$\frac{M}{m_p}\sum_g \frac{w_g^2}{p_g}\widehat{s}^2_{p\,yg}, \quad \frac{M}{m_p}\sum_g w_g \widehat{s}^2_{p\,yg}. \tag{11.72}$$

Infine, si stimano i valori ottimi di n_{sum} e m con:

$$\widehat{n}^*_{sum} = \sqrt{\frac{\frac{M}{m_p}\sum_g \frac{w_g^2}{p_g}\widehat{s}^2_{p\,yg}}{\frac{1}{M}\widehat{s}^2_{p\,b} - \frac{M}{m_p N}\sum_g w_g \widehat{s}^2_{p\,yg}}}\sqrt{\frac{Mc_1 + Nc_l}{c_2}}; \tag{11.73}$$

$$\widehat{m}^* = \frac{C_0}{c_1 + \frac{\widehat{n}^*_{sum}}{M}c_2 + \frac{N}{M}c_l}. \tag{11.74}$$

Sia il numero m_p di grappoli del campione pilota che i corrispondenti campioni di unità elementari dovrebbero essere "piccoli", ma tuttavia in grado di fornire stime sufficientemente accurate in (11.73), (11.74).

11.8 Campionamento a due stadi con stratificazione delle unità primarie*

In rilevazioni statistiche concrete, come già anticipato nella discussione sull'effetto del disegno, il disegno a due stadi è spesso combinato con altri disegni campionari, come quello stratificato. Un importante esempio è quello delle forze di lavoro, in cui il ruolo di grappoli (unità primarie) è svolto dai comuni. Questi sono preventivamente raggruppati in strati (formati su base sia demografica che geografica). Da ciascuno strato vengono in primo luogo selezionati alcuni comuni (in genere con un disegno non di tipo ssr); in secondo luogo, da ciascun comune selezionato è estratto un campione di famiglie, dei cui membri viene rilevato lo *status* occupazionale. Un disegno di questo tipo è null'altro che un esempio di piano di campionamento a due stadi, in cui i grappoli sono suddivisi in strati. In ciascuno strato si effettua, in modo indipendente, un campionamento a due stadi, cui al primo stadio sono selezionati grappoli, e al secondo unità elementari.

Formalmente, si consideri una popolazione suddivisa in H strati. Nel generico strato h vi sono M_h grappoli ($h = 1, \ldots, H$). A sua volta, il grappolo $g$mo dello strato $h$mo è formato da N_{hg} unità elementari ($g = 1, \ldots, M_h$; $h = 1, \ldots, H$). Nel seguito si adotterà la seguente simbologia, un po' pesante ma necessaria:

- $N_{h\cdot} = \sum_{g=1}^{M_h} N_{hg}$: numero di unità elementari dello strato h ($h = 1, \ldots, H$);
- $w_{hg} = \frac{N_{hg}}{N_{h\cdot}}$: peso del grappolo g nello strato h ($g = 1, \ldots, M_h$; $h = 1, \ldots, H$);
- $N = \sum_{h=1}^{H} N_{h\cdot} = \sum_{h=1}^{H} \sum_{g=1}^{M_h} N_{hg}$: numero totale di unità elementari della popolazione;
- $\omega_h = \frac{N_{h\cdot}}{N}$: peso dello strato h nella popolazione ($h = 1, \ldots, H$).

È immediato verificare che valgono le seguenti relazioni

$$\sum_{g=1}^{M_h} w_{hg} = 1 \ \textit{per ogni}\ h = 1, \ldots, H;$$

$$\sum_{h=1}^{H} \omega_h = 1.$$

Indichiamo poi y_{hgi} la modalità che il carattere di interesse $\mathcal{Y}$ assume in corrispondenza dell'unità elementare i del grappolo g dello strato h ($i = 1, \ldots, N_{hg}$; $g = 1, \ldots, M_h$; $h = 1, \ldots H$), e con:

- $\mu_{yhg} = \frac{1}{N_{hg}} \sum_{i=1}^{N_{hg}} y_{hgi}$: media del grappolo g dello strato h ($g = 1, \ldots, M_h$; $h = 1, \ldots H$);
- $S^2_{yhg} = \frac{1}{N_{hg}-1} \sum_{i=1}^{N_{hg}} (y_{hgi} - \mu_{yhg})^2$: varianza corretta del grappolo g dello strato h ($g = 1, \ldots, M_h$; $h = 1, \ldots H$);

- $\mu_{yh\cdot} = \frac{1}{N_{h\cdot}} \sum_{g=1}^{M_h} \sum_{i=1}^{N_{hg}} y_{hgi}$: media dello strato h ($h = 1, \ldots, H$);
- $\mu_y = \frac{1}{N} \sum_{h=1}^{H} \sum_{g=1}^{M_h} \sum_{i=1}^{N_{hg}} y_{hgi}$ media della popolazione.

Tra le varie medie sopra definite sussistono alcune relazioni fondamentali, di seguito riportate:

$$\mu_{yh\cdot} = \frac{1}{N_{h\cdot}} \sum_{g=1}^{M_h} \sum_{i=1}^{N_{hg}} y_{hgi} = \sum_{g=1}^{M_h} \frac{N_{hg}}{N_{h\cdot}} \left\{ \frac{1}{N_{hg}} \sum_{i=1}^{N_{hg}} y_{hgi} \right\} = \sum_{g=1}^{M_h} w_{hg} \mu_{yhg}; \quad (11.75)$$

$$\mu_y = \frac{1}{N} \sum_{h=1}^{H} \sum_{g=1}^{M_h} \sum_{i=1}^{N_{hg}} y_{hgi} = \sum_{h=1}^{H} \frac{N_{h\cdot}}{N} \left\{ \sum_{g=1}^{M_h} \frac{N_{hg}}{N_{h\cdot}} \left(\frac{1}{N_{hg}} \sum_{i=1}^{N_{hg}} y_{hgi} \right) \right\}$$

$$= \sum_{h=1}^{H} \omega_h \left\{ \sum_{g=1}^{M_h} w_{hg} \mu_{yhg} \right\} = \sum_{h=1}^{H} \omega_h \mu_{yh\cdot} \quad (11.76)$$

La (11.75) mostra che la media del grappolo $h$mo è esprimibile come media delle medie dei grappoli che lo compongono. La (11.76) mostra che la media della popolazione è esprimibile come media delle medie degli strati in cui è suddivisa.

Il disegno campionario a due stadi (semplici) con stratificazione delle unità primarie si realizza in modo molto semplice, qui di seguito descritto.

- *I stadio.* Da ciascuno stato h si seleziona, in modo indipendente, un campione ssr $\mathbf{g}_{m_h}$ di m_h grappoli ($h = 1, \ldots, H$).
- *II stadio.* Da ogni grappolo $g \in \mathbf{g}_{m_h}$ scelto al primo stadio in ciascuno degli strati si seleziona, mediante disegno ssr, un campione $\mathbf{s}_{hg}$ di n_{hg} unità elementari. I campioni $\mathbf{s}_{hg}$ di unità elementari appartenenti a strati diversi sono *indipendenti*. Inoltre, campioni di unità elementari di grappoli di uno stesso strato h sono *indipendenti dato* $\mathbf{g}_{m_h}$.

Sia ora

$$\overline{y}_{hg} = \frac{1}{n_{hg}} \sum_{i \in \mathbf{s}_{hg}} y_{hgi}$$

la media campionaria del grappolo g dello strato h. Come immediata estensione di quanto svolto nelle precedenti sezioni, come stimatore (corretto) della media μ_{yh} dello strato h si può fare riferimento a

$$\widehat{\mu}_{2sth} = \frac{1}{m_h} \sum_{g \in \mathbf{g}_{m_h}} M_h w_{hg} \overline{y}_{hg} \quad (11.77)$$

così che come stimatore corretto della media μ_y della popolazione si ha:

$$\widehat{\mu}_{2st;str} = \sum_{h=1}^{H} \omega_h \, \widehat{\mu}_{2sth}. \quad (11.78)$$

Per quanto riguarda la varianza di (11.78), è immediato verificare (Esercizio 11.9) che

$$\begin{aligned} & V(\widehat{\mu}_{2st;str}) \\ &= \sum_{h=1}^{H} \omega_h^2 V(\widehat{\mu}_{2sth}) \\ &= \sum_{h=1}^{H} \omega_h^2 \left\{ \left(\frac{1}{m_h} - \frac{1}{M_h} \right) S_{bh}^2 + \frac{M_h}{m_h} \sum_{g=1}^{M_h} w_{hg}^2 \left(\frac{1}{n_{hg}} - \frac{1}{N_{hg}} \right) S_{yhg}^2 \right\} \end{aligned} \quad (11.79)$$

essendo

$$S_{bh}^2 = \frac{1}{M_h - 1} \sum_{g=1}^{M_h} \left(M_h w_{hg} \mu_{yhg} - \mu_{yh\cdot} \right)^2 . \quad (11.80)$$

La stima della varianza (11.80), infine, non presenta difficoltà di rilievo. Un suo stimatore corretto, immediata estensione di quanto svolto nelle sezioni precedenti, è proposto nell'Esercizio 11.10.

Esercizi

11.1. Verificare che valgono le (11.23), (11.23).

11.2. Verificare che nel caso in cui (*i*) i grappoli hanno tutti la stessa numerosità L, e (*ii*) da ciascun grappolo vengono selezionate l unità elementari, lo stimatore (11.18) si riduce a (11.31).

11.3. Verificare che valgono le (11.34), (11.35).

11.4. Verificare che nel caso di disegno a due stadi semplici con numerosità costante, lo stimatore (11.18) è uno stimatore corretto di (11.11).

11.5. Provare che il problema di minimo vincolato (11.57) ha come soluzione (11.58).

Suggerimento. Posto per semplicità $a = S_b^2 - S_w^2/L$ e $b = S_w^2$, la funzione Lagrangiana assume la forma $\mathcal{L} = (a+b/l)/m + \lambda(m(c_1+c_2 l) - C_0)$. Da $\partial\mathcal{L}/\partial l = -b(ml^2) + \lambda m c_2 = 0$, $\partial\mathcal{L}/\partial m = -(a + b/l)/m^2 + \lambda(c_1 + c_2 l) = 0$ si ottengono rispettivamente le due equazioni $\lambda m^2 = b/(c_2 l^2)$, $\lambda m^2 = (a+b/l)/(c1+c_2 l)$, chie forniscono il valore ottimo di l. Dal vincolo si ottiene poi il valore ottimo di m.

11.6. Provare la relazione (11.63).

11.7. Provare che il problema di minimo vincolato (11.65) ha come soluzione (11.70), (11.71).

Suggerimento. Usare gli stessi argomenti dell'Esercizio 11.4.

11.8. Verificare la relazione (11.36).

11.9. Verificare la relazione (11.79).

11.10. Detta

$$\widehat{s}^2_{yhg} = \frac{1}{n_{hg}-1} \sum_{i \in \mathbf{s}_{hg}} (y_{hgi} - \overline{y}_{hg})^2$$

la varianza campionaria corretta del grappolo g dello strato h, si definiscano le seguenti quantità:

$$\widehat{V}_{1h} = \frac{1}{m_h - 1} \sum_{g \in \mathbf{g}_{m_h}} (M_h w_{hg} \overline{y}_{hg} - \widehat{\mu}_{2sth})^2;$$

$$\widehat{V}_{2h} = \frac{1}{m_h} \sum_{g \in \mathbf{g}_{m_h}} w^2_{hg} \left(\frac{1}{n_{hg}} - \frac{1}{N_{hg}} \right) \widehat{s}^2_{yhg};$$

$$\widehat{V}_{2sth} = \left(\frac{1}{m_h} - \frac{1}{M_h} \right) \widehat{V}_{1h} + M_h \widehat{V}_{2h}.$$

a. Verificare che $\widehat{V}_{2sth}$ è uno stimatore corretto di $V(\widehat{\mu}_{2sth})$.
b. Verificare che

$$\widehat{V}_{2st;str} = \sum_{h=1}^{H} \omega^2_h \, \widehat{V}_{2sth}$$

è uno stimatore corretto della (11.79).

12

Disegni campionari a probabilità variabile

12.1 Aspetti generali. Probabilità di inclusione

La nozione generale di disegno campionario, così come quelle di numerosità e di numerosità effettiva, sono già state fornite nel Capitolo 2. I disegni di campionamento visti fino ad ora, pur essendo composti da campioni che non hanno tutti necessariamente la stessa probabilità di selezione, hanno una caratteristica comune: sono *riconducibili al disegno semplice senza ripetizione* (ssr), nel senso che sono sostanzialmente suoi "derivati". In effetti, è evidente come i disegni di campionamento stratificato, a grappolo, a due stadi semplici, sono essenzialmente varianti del disegno ssr. In questo capitolo e nei successivi, invece, ci si concentrerà su disegni campionari *non riconducibili* a quello semplice, che verranno per brevità denominati *disegni a probabilità variabile*. Essi presentano importanti problemi di implementazione, in generale di non agevole soluzione. D'altra parte, disegni campionari a probabilità variabile sono ampiamente usati nella pratica applicativa.

Per lo studio di disegni campionari a probabilità variabile riveste importanza fondamentale la nozione di probabilità di inclusione. Consideriamo un generico disegno campionario $(\mathcal{S}, p(\cdot))$. La *probabilità di inclusione* (del primo ordine) *dell'unità* i è la probabilità di selezionare un campione $\mathbf{s} \in \mathcal{S}$ contenente l'unità i $(= 1, \ldots, N)$. L'inclusione di una data unità i nel campione $\mathbf{s}$ può essere espressa attraverso il ricorso alla variabile indicatrice

$$\delta(i; \mathbf{s}) = \begin{cases} 1 \text{ se } i \in \mathbf{s} \\ 0 \text{ se } i \notin \mathbf{s} \end{cases} \tag{12.1}$$

la quale assume il valore 1 se il campione $\mathbf{s}$ contiene l'unità i *almeno una volta*, e il valore 0 altrimenti. Si osservi che ciò implica che l'indicatore (12.1) *non dipende dal numero di volte in cui il campione contiene l'unità* i. L'unico elemento che determina il valore di $\delta(i; \mathbf{s})$ è l'appartenenza o meno dell'unità i al campione $\mathbf{s}$.

Tenendo conto che il campione $\mathbf{s}$ contiene l'unità i *se e solo se* $\delta(i; \mathbf{s}) = 1$, la probabilità di inclusione dell'unità i è null'altro che la probabilità che la

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_12,

v.a. $\delta(i;\,\mathbf{s})$ assuma il valore 1. In simboli, se si indica con π_i la probabilità di inclusione del primo ordine dell'unità i, si ha

$$\pi_i = Pr(\delta(i;\,\mathbf{s}) = 1). \tag{12.2}$$

Chiaramente, dalla (12.2) si desume che la probabilità di inclusione π_i si ottiene sommando le probabilità di tutti i campioni dello spazio campionario che contengono l'unità i. In simboli, detto $\mathcal{S}_i = \{\mathbf{s} \in \mathcal{S} : \mathbf{s} \ni i\}$ l'insieme dei campioni in $\mathcal{S}$ che contengono l'unità i, si può scrivere

$$\pi_i = \sum_{\mathbf{s}\in\mathcal{S}_i} p(\mathbf{s}). \tag{12.3}$$

In modo analogo è possibile definire la probabilità che due unità distinte i, j siano incluse nel campione, denominata probabilità di inclusione del secondo ordine e indicata con π_{ij}. Anche la probabilità π_{ij} può essere espressa in termini delle variabili indicatrici. Infatti, il campione $\mathbf{s}$ contiene le due unità distinte i, j se e solo se si ha simultaneamente $\delta(i;\,\mathbf{s}) = 1$ e $\delta(j;\,\mathbf{s}) = 1$. D'altra parte, è facile vedere che

$$\delta(i;\,\mathbf{s}) = 1 \textit{ e } \delta(j;\,\mathbf{s}) = 1 \;\; \textit{se e solo se } \delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s}) = 1$$

ossia il campione $\mathbf{s}$ contiene la coppia i, j di unità distinte se e solo se il prodotto dei due indicatori $\delta(i;\,\mathbf{s})$ e $\delta(j;\,\mathbf{s})$ è pari a 1. Pertanto, si può scrivere

$$\pi_{ij} = Pr(\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s}) = 1). \tag{12.4}$$

Anche in questo caso vale una relazione simile alla (12.3). Detto $\mathcal{S}_{ij}$ il sottoinsieme dello spazio campionario costituito da tutti i campioni che contengono le unità (i, j) ($\mathcal{S}_{ij} = \{\mathbf{s} \in \mathcal{S} : \mathbf{s} \ni (i, j)\}$), si ha:

$$\pi_{ij} = \sum_{\mathbf{s}\in\mathcal{S}_{ij}} p(\mathbf{s}). \tag{12.5}$$

È immediato verificare dalla (12.4) che vale la relazione (di *simmetria*) $\pi_{ij} = \pi_{ji}$.

La nozione di probabilità di inclusione del secondo ordine si può anche estendere a coppie di unità coincidenti. In effetti, se $j = i$ si ha $\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s}) = \delta(i;\,\mathbf{s})^2 = \delta(i;\,\mathbf{s})$ (perché $\delta(i;\,\mathbf{s})$ assume i valori 1 o 0, che elevati al quadrato sono ancora rispettivamente pari a 1 e 0); di conseguenza

$$\pi_{ii} = Pr(\delta(i;\,\mathbf{s})^2 = 1) = Pr(\delta(i;\,\mathbf{s}) = 1) = \pi_i. \tag{12.6}$$

Esempio 12.1. Riprendiamo l'Esempio 2.2. del Capitolo 2, in cui si considera una popolazione di $N = 7$ unità: $I_7 = \{1, 2, \ldots, 7\}$. Lo spazio dei campioni è formato dai sei campioni:

$$\mathbf{s}_1 = \{1, 2\},\ \mathbf{s}_2 = \{1, 3\},\ \mathbf{s}_3 = \{4\},\ \mathbf{s}_4 = \{2, 3, 5\},\ \mathbf{s}_5 = \{6, 7\},\ \mathbf{s}_6 = \{4, 5, 7\}$$

con le seguenti probabilità:

$p(\mathbf{s}_1) = 0.15,\ p(\mathbf{s}_2) = 0.2,\ p(\mathbf{s}_3) = 0.1,\ p(\mathbf{s}_4) = 0.1,\ p(\mathbf{s}_5) = 0.15,\ p(\mathbf{s}_6) = 0.3.$

Le probabilità di inclusione del primo ordine delle unità sono:

$$\pi_1 = \sum_{\mathbf{s}\in\mathcal{S}_1} p(\mathbf{s}) = p(\mathbf{s}_1) + p(\mathbf{s}_2) = 0.15 + 0.2 = 0.35$$

$$\pi_2 = \sum_{\mathbf{s}\in\mathcal{S}_2} p(\mathbf{s}) = p(\mathbf{s}_1) + p(\mathbf{s}_4) = 0.15 + 0.1 = 0.25$$

$$\pi_3 = \sum_{\mathbf{s}\in\mathcal{S}_3} p(\mathbf{s}) = p(\mathbf{s}_2) + p(\mathbf{s}_4) = 0.2 + 0.1 = 0.3$$

$$\pi_4 = \sum_{\mathbf{s}\in\mathcal{S}_4} p(\mathbf{s}) = p(\mathbf{s}_3) + p(\mathbf{s}_6) = 0.1 + 0.3 = 0.4$$

$$\pi_5 = \sum_{\mathbf{s}\in\mathcal{S}_5} p(\mathbf{s}) = p(\mathbf{s}_4) + p(\mathbf{s}_6) = 0.1 + 0.3 = 0.4$$

$$\pi_6 = \sum_{\mathbf{s}\in\mathcal{S}_6} p(\mathbf{s}) = p(\mathbf{s}_5) = 0.15$$

$$\pi_7 = \sum_{\mathbf{s}\in\mathcal{S}_7} p(\mathbf{s}) = p(\mathbf{s}_5) + p(\mathbf{s}_6) = 0.15 + 0.3 = 0.45.$$

In modo simile si possono calcolare le probabilità di inclusione del secondo ordine. □

La nozione di probabilità di inclusione è definita per disegni campionari del tutto generali, con o senza ripetizioni, ordinati o non ordinati. Tuttavia, ci si può ridurre a considerare solo disegni non ordinati e senza ripetizioni, in quanto è facile provare che la riduzione di un qualsiasi disegno campionario possiede le stesse probabilià di inclusione del disegno di partenza. Questo risultato, semplice ma importante, è provato nella successiva proposizione.

Proposizione 12.1. *Se $(\mathcal{S}, p(\cdot))$ è un disegno campionario, e se $(\mathcal{S}^*, p^*(\cdot))$ è la sua riduzione, i due disegni campionari hanno le stesse probabilità di inclusione.*

Dimostrazione. Ci limitiamo per brevità alle sole probabilità di inclusione del primo ordine, in quanto per quelle del secondo ordine vale un discorso praticamente identico. La notazione è identica a quella del Capitolo 2. Siano π_i, π_i^* le probabilità di inclusione dell'unità i rispettivamente nel disegno "originario" e in quello ridotto. Preso un qualunque $\mathbf{s}^* \in \mathcal{S}^*$, sia inoltre $C(\mathbf{s}^*) = \{\mathbf{s} \in \mathcal{S} :\ r(\mathbf{s}) = \mathbf{s}^*\}$ l'insieme dei campioni "originari" aventi $\mathbf{s}^*$ come riduzione. Infine, siano $\mathcal{S}_i$ e $\mathcal{S}_i^*$ rispettivamente l'insieme dei campioni "originari" e di quelli "ridotti" contenenti l'unità i. Chiaramente, vale la relazione

$$\mathcal{S}_i^* = \{r(\mathbf{s});\ \mathbf{s} \in \mathcal{S}_i\} \tag{12.7}$$

dalla quale discende che

$$\begin{aligned}\pi_i^* &= \sum_{\mathbf{s}^* \in \mathcal{S}_i^*} p(\mathbf{s}^*) \\ &= \sum_{\mathbf{s}^* \in \mathcal{S}_i^*} \left\{ \sum_{\mathbf{s} \in C(\mathbf{s}^*)} p(\mathbf{s}) \right\} \\ &= \sum_{\mathbf{s} \in \mathcal{S}_i} p(\mathbf{s}) \\ &= \pi_i. \end{aligned}$$

□

Esempio 12.2. Consideriamo il disegno campionario dell'Esempio 2.1, in cui si ha una popolazione di $N = 7$ unità: $I_7 = \{1, 2, \ldots, 7\}$. Lo spazio dei campioni sia formato dagli otto campioni:

$$\begin{aligned}&\mathbf{s}_1 = (1, 2, 3),\ \mathbf{s}_2 = (1, 2, 4),\ \mathbf{s}_3 = (5, 6),\ \mathbf{s}_4 = (7), \\ &\mathbf{s}_5 = (6, 5),\ \mathbf{s}_6 = (1, 2, 2, 3),\ \mathbf{s}_7 = (3, 1, 2),\ \mathbf{s}_8 = (3, 1, 1, 2)\end{aligned}$$

con le seguenti probabilità

$$\begin{aligned}&p(\mathbf{s}_1) = 0.1,\ p(\mathbf{s}_2) = 0.15,\ p(\mathbf{s}_3) = 0.15,\ p(\mathbf{s}_4) = 0.05, \\ &p(\mathbf{s}_5) = 0.2,\ p(\mathbf{s}_6) = 0.05,\ p(\mathbf{s}_7) = 0.1,\ p(\mathbf{s}_8) = 0.2.\end{aligned}$$

Le probabilità di inclusione del primo ordine sono uguali a

$$\begin{aligned}\pi_1 &= = p(\mathbf{s}_1) + p(\mathbf{s}_2) + p(\mathbf{s}_6) + p(\mathbf{s}_7) + p(\mathbf{s}_8) = 0.1 + 0.15 + 0.05 + 0.1 + 0.2 \\ &= 0.6 \\ \pi_2 &= p(\mathbf{s}_1) + p(\mathbf{s}_2) + p(\mathbf{s}_6) + p(\mathbf{s}_7) + p(\mathbf{s}_8) = 0.1 + 0.15 + 0.05 + 0.1 + 0.2 \\ &= 0.6 \\ \pi_3 &= p(\mathbf{s}_1) + p(\mathbf{s}_6) + p(\mathbf{s}_7) + p(\mathbf{s}_8) = 0.1 + 0.05 + 0.1 + 0.2 = 0.45 \\ \pi_4 &= p(\mathbf{s}_2) = 0.15 \\ \pi_5 &= p(\mathbf{s}_3) + p(\mathbf{s}_5) = 0.15 + 0.2 = 0.35 \\ \pi_6 &= p(\mathbf{s}_3) + p(\mathbf{s}_5) = 0.15 + 0.2 = 0.35 \\ \pi_7 &= p(\mathbf{s}_4) = 0.05.\end{aligned}$$

La riduzione di questo disegno campionario è stata costruita nell'Esempio 2.5. Lo spazio dei campioni ridotto $\mathcal{S}^*$ è formato dai campioni:

$$\mathbf{s}_1^* = \{1, 2, 3\},\ \mathbf{s}_2^* = \{1, 2, 4\},\ \mathbf{s}_3^* = \{5, 6\},\ \mathbf{s}_4^* = \{7\}$$

con probabilità

$$p^*(\mathbf{s}_1^*) = 0.45,\ p^*(\mathbf{s}_2^*) = 0.15,\ p^*(\mathbf{s}_3^*) = 0.35,\ p(\mathbf{s}_4^*) = 0.05.$$

Le probabilità di inclusione calcolate in base al disegno campionario ridotto sono pari a

$$\begin{aligned}
\pi_1^* &= p^*(\mathbf{s}_1^*) + p^*(\mathbf{s}_2^*) = 0.45 + 0.15 = 0.6 \\
\pi_2^* &= p^*(\mathbf{s}_1^*) + p^*(\mathbf{s}_2^*) = 0.45 + 0.15 = 0.6 \\
\pi_3^* &= p^*(\mathbf{s}_1^*) = 0.45 \\
\pi_4^* &= p^*(\mathbf{s}_2^*) = 0.15 \\
\pi_5^* &= p^*(\mathbf{s}_3^*) = 0.35 \\
\pi_6^* &= p^*(\mathbf{s}_3^*) = 0.35 \\
\pi_7^* &= p^*(\mathbf{s}_4^*) = 0.05
\end{aligned}$$

e chiaramente coincidono con le π_i. □

In forza della Proposizione 12.1 nel seguito si assumerà sempre, senza perdita di generalità, che il disegno di campionamento sia non ordinato e senza ripetizioni.

In generale è possibile definire anche probabilità di inclusione di ordine superiore al secondo ma poiché queste rivestono un ruolo meno importante nell'ambito dell'inferenza da popolazioni finite il loro calcolo verrà tralasciato.

Le proprietà delle variabili indicatrici $\delta(i;\, \mathbf{s})$ sono riassunte nella seguente proposizione.

Proposizione 12.2. *Dato un generico disegno di campionamento* $(\mathcal{S},\, p(\cdot))$*, e per* $i = 1, \ldots, N$ *valgono le seguenti proprietà*

$$E[\delta(i;\, \mathbf{s})] = \pi_i; \tag{12.8}$$

$$V[\delta(i;\, \mathbf{s})] = \pi_i(1 - \pi_i); \tag{12.9}$$

$$E[\delta(i;\, \mathbf{s})\, \delta(j;\, \mathbf{s})] = \pi_{ij}; \tag{12.10}$$

$$C[\delta(i;\, \mathbf{s}),\, \delta(j;\, \mathbf{s})] = \pi_{ij} - \pi_i \pi_j. \tag{12.11}$$

Dimostrazione. Per provare (12.8), (12.9) è sufficiente osservare che $\delta(i;\, \mathbf{s})$, fissata l'unità i, è una v.a. di Bernoulli, che assume i due valori 1, 0 rispettivamente con probabilità π_i e $1 - \pi_i$. Si ha quindi

$$E[\delta(i;\, \mathbf{s})] = 1 \times Pr(\delta(i;\, \mathbf{s}) = 1) + 0 \times Pr(\delta(i;\, \mathbf{s}) = 0) = \pi_i$$

e analogamente

$$\begin{aligned}
V[\delta(i;\, \mathbf{s})] &= E[\delta(i;\, \mathbf{s})^2] - \{E[\delta(i;\, \mathbf{s})]\}^2 \\
&= E[\delta(i;\, \mathbf{s})] - \pi_i^2 \\
&= \pi_i - \pi_i^2 \\
&= \pi_i(1 - \pi_i).
\end{aligned}$$

Per quanto riguarda le (12.10), (12.11), notiamo che il prodotto $\delta(i;\mathbf{s})\delta(j;\mathbf{s})$ definisce una nuova variabile indicatrice che assume valore pari a uno se e solo se entrambe le unità i e j sono incluse nel campione, ciò implica che $E[\delta(i;\mathbf{s})\delta(j;\mathbf{s})] = \pi_{ij}$. Da tale risultato discende che

$$\begin{aligned} E[\delta(i;\mathbf{s})\,\delta(j;\mathbf{s})] &= \pi_{ij} \\ C[\delta(i;\mathbf{s}),\delta(j;\mathbf{s})] &= E[\delta(i;\mathbf{s})\delta(j;\mathbf{s})] - E[\delta(i;\mathbf{s})]E[\delta(j;\mathbf{s})] \\ &= \pi_{ij} - \pi_i\pi_j. \end{aligned}$$

Chiaramente il segno di tali covarianze dipenderà dalle caratteristiche del disegno campionario. □

I risultati della Proposizione 12.2 possono anche essere posti in forma vettoriale. Definiamo il vettore di N elementi (uno per ogni unità della popolazione)

$$\boldsymbol{\delta}(\mathbf{s}) = \begin{bmatrix} \delta(1;\mathbf{s}) \\ \delta(2;\mathbf{s}) \\ \cdots \\ \delta(N;\mathbf{s}) \end{bmatrix} \tag{12.12}$$

in cui ciascuna componente è l'indicatore di presenza-assenza della corrispondente unità nel campione $\mathbf{s}$. La conoscenza del vettore $\boldsymbol{\delta}(\mathbf{s})$ *equivale* alla conoscenza del campione $\mathbf{s}$. Le componenti di $\boldsymbol{\delta}(\mathbf{s})$ pari a 1 corrispondono alle unità presenti nel campione $\mathbf{s}$, mentre le componenti pari a 0 corrispondono alle unità non presenti in $\mathbf{s}$.

Indichiamo con

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \cdots \\ \pi_N \end{bmatrix} \tag{12.13}$$

il vettore delle probabilità di inclusione del primo ordine, e con

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1N} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{N1} & \pi_{N2} & \cdots & \pi_{NN} \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_{12} & \cdots & \pi_{1N} \\ \pi_{12} & \pi_2 & \cdots & \pi_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{1N} & \pi_{2N} & \cdots & \pi_N \end{bmatrix} \tag{12.14}$$

la matrice delle probabilità di inclusione del secondo ordine.

Come conseguenza della Proposizione 12.2 il valore del vettore $\boldsymbol{\delta}(\mathbf{s})$ è pari a $\boldsymbol{\pi}$:

$$E[\boldsymbol{\delta}(\mathbf{s})] = \boldsymbol{\pi}. \tag{12.15}$$

Indicando poi con $\boldsymbol{\delta}(\mathbf{s})'$ il trasposto del vettore $\boldsymbol{\delta}(\mathbf{s})$, la (12.10) si può scrivere in forma matriciale come

$$E[\boldsymbol{\delta}(\mathbf{s})\boldsymbol{\delta}(\mathbf{s})'] = \Pi. \tag{12.16}$$

Infine, la (12.11) si può esprimere in forma matriciale poiché la matrice di varianze e covarianze di $\boldsymbol{\delta}(\mathbf{s})$ è pari a $\boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}'$. In simboli:

$$E[(\boldsymbol{\delta}(\mathbf{s}) - E[\boldsymbol{\delta}(\mathbf{s})])\,(\boldsymbol{\delta}(\mathbf{s}) - E[\boldsymbol{\delta}(\mathbf{s})])'] = \boldsymbol{\Pi} - \boldsymbol{\pi}\boldsymbol{\pi}'. \tag{12.17}$$

Nel prosieguo, se non diversamente specificato, considereremo disegni campionari con probabilità di inclusione del primo ordine strettamente positive. Formalmente

$$\pi_i > 0 \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{12.18}$$

Ciò significa che ogni elemento della popolazione ha la possibilità di entrare a far parte del campione. Notiamo che se oltre alla condizione (12.18) il disegno possiede anche probabilità di inclusione del secondo ordine strettamente positive, formalmente

$$\pi_{ij} > 0 \quad \forall i, j \in U (i \neq j)$$

il disegno è detto *misurabile* . Come vedremo la nozione di misurabilità di un disegno riveste una notevole importanza per l'esistenza di stimatori corretti della varianza.

12.2 Proprietà delle probabilità di inclusione

Consideriamo un generico disegno di campionamento $(\mathcal{S}, p(\cdot))$. L'*ampiezza media* di un disegno campionario, indicata con $\overline{n}$, è il numero medio di unità contenute nei campioni. In simboli:

$$\overline{n} = E[n(\mathbf{s})] = \sum_{\mathbf{s}\in\mathcal{S}} n(\mathbf{s})\, p(\mathbf{s}).$$

Similmente, l'*ampiezza media effettiva* di un disegno campionario, indicata con $\overline{\nu}$, è il numero medio di unità *distinte* contenute nei campioni:

$$\overline{\nu} = E[\nu(\mathbf{s})] = \sum_{\mathbf{s}\in\mathcal{S}} \nu(\mathbf{s})\, p(\mathbf{s}).$$

Chiaramente, è sempre $\overline{\nu} \leqslant \overline{n}$. Inoltre, l'uguaglianza $\overline{\nu} = \overline{n}$ vale se e solo se il disegno campionario è senza ripetizioni. Le relazioni che legano l'ampiezza media effettiva di un disegno alle probabilità di inclusione sono illustrate nella seguente proposizione.

Proposizione 12.3. *Dato un generico disegno di campionamento* $(\mathcal{S}, p(\cdot))$*, le probabilità di inclusione del primo e secondo ordine soddisfano le seguenti proprietà*

$$\sum_{i=1}^{N} \pi_i = \overline{\nu}; \tag{12.19}$$

$$\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\pi_{ij} = V[\nu(\mathbf{s})] + \overline{\nu}(\overline{\nu}-1); \tag{12.20}$$

dove $V[\nu(\mathbf{s})]$ *rappresenta la varianza della variabile aleatoria* $\nu(\mathbf{s})$.

Dimostrazione. Per provare la (12.19) è sufficiente osservare che

$$\sum_{i=1}^{N}\delta(i;\,\mathbf{s}) = \nu(\mathbf{s})\,. \tag{12.21}$$

da cui si ottiene subito

$$\sum_{i=1}^{N}\pi_i = \sum_{i=1}^{N}E[\delta(i;\,\mathbf{s})] = E\left[\sum_{i=1}^{N}\delta(i;\,\mathbf{s})\right] = E[\nu(\mathbf{s})]\,.$$

Per provare la (12.20) basta usare la relazione (conseguenza della (12.21))

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s}) &= \sum_{i=1}^{N}\delta(i;\,\mathbf{s})\left\{\sum_{\substack{j=1\\ j\neq i}}^{N}\delta(j;\,\mathbf{s})\right\}\\
&= \sum_{i=1}^{N}\delta(i;\,\mathbf{s})\left\{\sum_{j=1}^{N}\delta(j;\,\mathbf{s}) - \delta(i;\,\mathbf{s})\right\}\\
&= \sum_{i=1}^{N}\delta(i;\,\mathbf{s})\,\{\nu(\mathbf{s}) - \delta(i;\,\mathbf{s})\}\\
&= \nu(\mathbf{s})\sum_{i=1}^{N}\delta(i;\,\mathbf{s}) - \sum_{i=1}^{N}\delta(i;\,\mathbf{s})^2\\
&= \nu(\mathbf{s})^2 - \nu(\mathbf{s})
\end{aligned} \tag{12.22}$$

dalla quale si ottiene

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\pi_{ij} &= \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}E[\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s})]\\
&= E\left[\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\delta(i;\,\mathbf{s})\,\delta(j;\,\mathbf{s})\right]\\
&= E[\nu(\mathbf{s})^2] - E[\nu(\mathbf{s})]\\
&= V[\nu(\mathbf{s})] + \overline{\nu}(\overline{\nu}-1).
\end{aligned}$$

□

Se il disegno campionario è ad *ampiezza effettiva costante*, formalmente $\nu(\mathbf{s}) = \nu$ per ogni $\mathbf{s} \in \mathcal{S}$ (e quindi anche $\overline{\nu} = \nu$), allora vale anche la seguente ulteriore proprietà.

Proposizione 12.4. *Dato un generico disegno di campionamento* $(\mathcal{S}, p(\cdot))$ *ad ampiezza effettiva costante* ν*, le probabilità di inclusione del secondo ordine soddisfano la seguente relazione:*

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = (\nu - 1)\pi_i;\ i = 1, \ldots, N. \tag{12.23}$$

Dimostrazione. Usando le stesse considerazioni che portano alla (12.22), osserviamo in primo luogo che se tutti i campioni hanno lo stesso numero ν di unità distinte si ha

$$\begin{aligned}
\sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(i;\, \mathbf{s})\, \delta(j;\, \mathbf{s}) &= \delta(i;\, \mathbf{s}) \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(j;\, \mathbf{s}) \\
&= \delta(i;\, \mathbf{s}) \left\{ \sum_{j=1}^{N} \delta(j;\, \mathbf{s}) - \delta(i;\, \mathbf{s}) \right\} \\
&= \delta(i;\, \mathbf{s}) \{\nu - \delta(i;\, \mathbf{s})\} \\
&= \nu\, \delta(i;\, \mathbf{s}) - \delta(i;\, \mathbf{s}) \\
&= (\nu - 1)\delta(i;\, \mathbf{s})
\end{aligned} \tag{12.24}$$

da cui si ottiene

$$\begin{aligned}
\sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} &= \sum_{\substack{j=1 \\ j \neq i}}^{N} E[\delta(i;\, \mathbf{s})\, \delta(j;\, \mathbf{s})] \\
&= E\left[\sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(i;\, \mathbf{s})\, \delta(j;\, \mathbf{s}) \right] \\
&= E[(\nu - 1)\delta(i;\, \mathbf{s})] \\
&= (\nu - 1)\pi_i\,.
\end{aligned}$$

□

Se il disegno di campionamento è senza ripetizioni allora le proposizioni 12.3 e 12.4 si applicano all'ampiezza media $\overline{n}$. Inoltre nel caso in cui il disegno sia anche ad ampiezza costante vale il seguente risultato.

Proposizione 12.5. *Se il disegno campionario è senza ripetizioni* ($\nu(\mathbf{s}) = n(\mathbf{s})$ *per ogni* $\mathbf{s} \in \mathcal{S}$) *e ad ampiezza costante* ($n(\mathbf{s}) = n$ *per ogni* $\mathbf{s} \in \mathcal{S}$) *allora valgono le seguenti relazioni*

$$\sum_{i=1}^{N} \pi_i = n; \quad \sum_{i=1}^{N} \sum_{\substack{j=i \\ j \neq i}}^{N} \pi_{ij} = n(n-1); \quad \sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = (n-1)\pi_i \,. \tag{12.25}$$

Dimostrazione. È una conseguenza delle proposizioni 12.3, 12.4. □

12.3 Probabilità di inclusione per disegni campionari "semplici"

In questa sezione vengono forniti esempi di calcolo delle probabilità di inclusione per alcuni disegni campionari particolarmente importanti, legati al disegno semplice.

Esempio 12.3 (Disegno semplice senza ripetizione). In un campionamento casuale semplice senza ripetizione (ssr) la probabilità di inclusione del primo ordine per una generica unità i della popolazione è data da

$$\pi_i = \sum_{\mathbf{s} \in \mathcal{S}_i} p(\mathbf{s}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \tag{12.26}$$

per $i = 1, \ldots, N$, essendo esattamente pari a $\binom{N-1}{n-1}$ i campioni $\mathbf{s}$ dello spazio campionario che includono l'unità i. Analogamente per le probabilità di inclusione del secondo ordine si ricava che

$$\pi_{ij} = \sum_{\mathbf{s} \in \mathcal{S}_{ij}} p(\mathbf{s}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} \tag{12.27}$$

essendo esattamente pari a $\binom{N-2}{n-2}$ i campioni dello spazio campionario che includono contemporaneamente le unità i e j (con $i \neq j$). Notiamo che nel campionamento semplice senza ripetizione tutte le unità della popolazione hanno la stessa probabilità di inclusione del primo ordine (pari alla frazione sondata). Se il disegno campionario presenta tale caratteristica si definisce *autoponderante*. □

Esempio 12.4 (Disegno semplice con ripetizione). In un campionamento casuale semplice con ripetizione (scr) la probabilità di inclusione del primo ordine per una generica unità i della popolazione è data da

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n. \tag{12.28}$$

Tale probabilità si calcola facilmente osservando che un campione non contiene l'unità i se e solo se è una disposizione senza ripetizione di classe n delle $N-1$ unità $I_N \setminus \{i\}$. Si ha quindi

$$\begin{aligned}\pi_i &= 1 - Pr(\mathbf{s} \not\ni i) \\ &= 1 - \frac{(N-1)^n}{N^n}\end{aligned}$$

da cui la (12.28).

In modo analogo si ricavano le probabilità di inclusione del secondo ordine per $i \neq j$:

$$\begin{aligned}\pi_{ij} &= 1 - Pr(\mathbf{s} \not\ni i) - Pr(\mathbf{s} \not\ni j) + Pr(\{\mathbf{s} \not\ni i\} \cap \{\mathbf{s} \not\ni j\}) \\ &= 1 - \frac{(N-1)^n}{N^n} - \frac{(N-1)^n}{N^n} + \frac{(N-2)^n}{N^n} \\ &= 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n. \end{aligned} \tag{12.29}$$

Inoltre sulla base della Proposizione 12.3, l'*ampiezza media effettiva* $\overline{\nu}$ di un disegno semplice con ripetizione risulta pari a

$$\begin{aligned}\overline{\nu} &= \sum_{i=1}^{N} \pi_i \\ &= \sum_{i=1}^{N} \left[1 - \left(1 - \frac{1}{N}\right)^n\right] \\ &= N\left[1 - \left(1 - \frac{1}{N}\right)^n\right]. \end{aligned} \tag{12.30}$$

□

Esempio 12.5 (Disegno campionario stratificato). In un disegno di campionamento stratificato le probabilità di inclusione si ricavano applicando ad ogni strato della popolazione i risultati ottenuti per il campionamento casuale semplice senza ripetizione (Esempio 12.3). Poiché l'unità i dello strato g è contenuta nel campione $\mathbf{s}$ se e solo se è contenuta nel "sottocampione" $\mathbf{s}_g$

dello strato g, la sua probabilità di inclusione del primo ordine è pari, con ovvia notazione, a

$$\begin{aligned}\pi_{(g)i} &= Pr(i \in \mathbf{s}_g) \\ &= \sum_{\mathbf{s}_g \in \mathcal{S}_{gi}} p(\mathbf{s}) \\ &= \frac{\binom{N_g - 1}{n_g - 1}}{\binom{N_g}{n_g}} \\ &= \frac{n_g}{N_g} \end{aligned} \tag{12.31}$$

con $\mathcal{S}_{gi} = \{\mathbf{s}_g \in \mathcal{C}_{N_g,n_g} : \mathbf{s}_g \ni i\}$.

Per il calcolo delle probabilità di inclusione del secondo ordine occorre distinguere a seconda che le unità i e j appartengano allo stesso strato g oppure a strati diversi, rispettivamente g, g'. Con lo stesso ragionamento usato per la probabilità di inclusione del primo ordine, nel primo caso si ricava che

$$\begin{aligned}\pi_{(g)ij} &= Pr((i,j) \in \mathbf{s}_g) \\ &= \sum_{\mathbf{s}_g \in \mathcal{S}_{gij}} p(\mathbf{s}) \\ &= \frac{\binom{N_g - 2}{n_g - 2}}{\binom{N_g}{n_g}} \\ &= \frac{n_g(n_g - 1)}{N_g(N_g - 1)} \end{aligned} \tag{12.32}$$

dove $\mathcal{S}_{gij} = \{\mathbf{s}_g \in \mathcal{C}_{N_g,n_g} : \mathbf{s}_g \ni (i,j)\}$. Se invece le unità appartengono a due strati distinti, essendo indipendenti i "sottocampioni" $\mathbf{s}_g$, $\mathbf{s}_{g'}$, si ricava che

$$\begin{aligned}\pi_{(gg')ij} &= Pr((i \in \mathbf{s}_g) \cap (j \in \mathbf{s}_{g'})) \\ &= Pr(i \in \mathbf{s}_g) Pr(j \in \mathbf{s}_{g'}) \\ &= \frac{n_g n_{g'}}{N_g N_{g'}}. \end{aligned} \tag{12.33}$$

□

Esempio 12.6 (Disegno campionario a grappolo). In un disegno campionario a grappolo la probabilità di inclusione per l'unità i del grappolo g è data, con notazione simile a quella usata nell'esempio precedente, da

$$\begin{aligned}\pi_{(g)i} &= Pr(g \in \mathbf{g}_m) \\ &= \frac{m}{M}. \end{aligned} \tag{12.34}$$

Infatti, poiché tutte le unità dei grappoli campionati entrano a far parte del campione, la probabilità di inclusione del primo ordine per l'unità i del grappolo g è uguale alla probabilità di inclusione del grappolo g.

Per il calcolo delle probabilità di inclusione del secondo ordine occorre distinguere a seconda che le unità i e j appartengano allo stesso grappolo oppure a grappoli diversi. Nel primo caso si ricava che

$$\begin{aligned} \pi_{(g)ij} &= Pr(g \in \mathbf{g}_m) \\ &= \frac{m}{M}\,. \end{aligned} \tag{12.35}$$

Analogamente se invece le unità i, j appartengono rispettivamente ai due grappoli distinti g e $g^{'}$, si ottiene

$$\begin{aligned} \pi_{(gg')ij} &= Pr((g, g^{'}) \in \mathbf{g}_m) \\ &= \frac{m(m-1)}{M(M-1)}. \end{aligned} \tag{12.36}$$

□

Esempio 12.7 (Disegno campionario a due stadi semplici). In un disegno campionario a due stadi semplici la probabilità di inclusione dell'unità i del grappolo g è data dalla

$$\begin{aligned} \pi_{(g)i} &= Pr(g \in \mathbf{g}_m) Pr(i \in s_g \,|\mathbf{g}_m) \\ &= \frac{m}{M}\frac{n_g}{N_g} \end{aligned} \tag{12.37}$$

in cui $Pr(g \in \mathbf{g}_m)$ rappresenta la probabilità di selezione del grappolo g (al primo stadio), mentre $Pr(i \in s_g \,|\mathbf{g}_m)$ è la probabilità che l'unità i appartenente al grappolo g sia estratta al secondo stadio di campionamento, essendo stato selezionato al primo stadio il grappolo g. Per quanto riguarda le probabilità di inclusione del secondo ordine, occorre distinguere a seconda che le unità i e j appartengano allo stesso grappolo g oppure a due diversi grappoli g, $g^{'}$. Nel primo caso si ricava che

$$\begin{aligned} \pi_{(g)ij} &= Pr(g \in \mathbf{g}_m)\, Pr((i,j) \in s_g \,|\mathbf{g}_m) \\ &= \frac{m}{M}\frac{n_g(n_g-1)}{N_g(N_g-1)}. \end{aligned} \tag{12.38}$$

Se invece le unità appartengono a grappoli distinti, rispettivamente g e $g^{'}$, si ottiene

$$\begin{aligned} \pi_{(gg')ij} &= Pr((g, g^{'}) \in \mathbf{g}_m) Pr(i \in s_g \,|\mathbf{g}_m) Pr(j \in s_{g'} \,|\mathbf{g}_m) \\ &= \frac{m(m-1)}{M(M-1)}\frac{n_g}{N_g}\frac{n_{g'}}{N_{g'}}. \end{aligned} \tag{12.39}$$

□

Esempio 12.8. Un disegno sistematico in cui i campioni hanno tutti la stessa numerosità n e il rapporto $M = N/n$ è un numero intero equivale ad un disegno a grappolo con M grappoli tutti della stessa numerosità n, e ciascuno formato dalle unità

$$\{g,\ g+M,\ g+2M,\ \ldots,\ g+(n-1)M\};\ \ g=1,\ \ldots,\ M.$$

Ciascuna unità ha probabilità di inclusione del primo ordine pari a $1/M = n/N$. Due unità distinte i, j hanno probabilità di inclusione del secondo ordine pari a n/N se appartengono allo stesso grappolo, e pari a 0 altrimenti. Osservando che due unità i, j appartengono allo stesso grappolo se e solo se $|i-j|$ è un multiplo di M, cioè se e solo se $|i-j|/M$ è intero, si può quindi scrivere

$$\pi_i = \frac{n}{N};\ \ \pi_{ij} = \begin{cases} \frac{n}{N} \textit{ se } \frac{|i-j|}{M} \textit{ è intero} \\ 0 \textit{ altrimenti} \end{cases}. \qquad \square$$

12.4 Estensioni immediate dei disegni campionari semplici: disegni *ppswr* e *ppswor*. Disegno di Midzuno-Lahiri

In generale, la scelta del disegno campionario va legata all'efficienza della risultante stima del parametro della popolazione, e da questo punto di vista non può essere disgiunta dal problema della scelta dello stimatore da utilizzare. Questo aspetto sarà evidente nei Capitoli 14, 15, soprattutto quando si studieranno disegni legati allo stimatore di Horvitz-Thompson. L'obiettivo di questa sezione è invece molto più limitato, e riguarda l'introduzione di alcuni disegni campionari che da un lato possono essere visti come immediate estensioni dei disegni semplici con e senza ripetizioni (disegni *ppswr* e *ppswor*), e dall'altro permettono di acquisire un minimo di familiarità con disegni di tipo "non semplice".

12.4.1 Disegno campionario ppswr

Supponiamo che per ciascuna unità i della popolazione I_N sia assegnato, in un qualche modo, un numero p_i positivo che ne misura l'"importanza" (*size*). Più grande p_i, più "importante" l'unità i. Per convenzione, e senza perdita di generalità, assumeremo che $p_1 + p_2 + \cdots + p_N = 1$.

Il disegno scr, introdotto nel Capitolo 3, si basa su un'idea elementare: si effettuano n prove indipendenti, in ciascuna delle quali si seleziona con uguale probabilità $1/N$ una delle unità della popolazione. L'idea-guida del disegno *ppswr* (*probability proportional to size with replacement*) è un'immediata estensione: si effettuano n prove indipendenti, in ciascuna delle quali

si seleziona una delle unità della popolazione; in ciascuna prova l'unità i ha probabilità p_i di essere selezionata.

Lo spazio dei campioni $\mathcal{S}$ è l'insieme di tutte le n-ple ordinate (disposizioni con ripetizione) $(i_1, i_2, \ldots, i_n)$ di n unità della popolazione. In simboli:

$$\mathcal{S} = \underbrace{I_N \times I_N \times \cdots \times I_N}_{n \text{ volte}} = I_N^n.$$

Un campione $\mathbf{s} = (i_1, i_2, \ldots, i_n)$ ha probabilità pari al prodotto delle probabilità di selezione delle singole unità che lo compongono:

$$p(\mathbf{s}) = p_{i_1} p_{i_2} \cdots p_{i_n}.$$

Com'è immediato constatare, il disegno scr corrisponde al caso speciale in cui $p_1 = \cdots = p_N = \frac{1}{N}$.

Con ragionamenti simili a quelli descritti per il disegno semplice con ripetizione si possono calcolare le probabilità di inclusione del primo e del secondo ordine. Si ha:

$$\pi_i = 1 - (1 - p_i)^n, \quad i = 1, \ldots, N; \tag{12.40}$$

$$\pi_{ij} = 1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n, \quad i \neq j = 1, \ldots, N. \tag{12.41}$$

12.4.2 Disegno campionario ppswor

Il punto di partenza del disegno *ppswor* (*probability proportional to size without replacement*) è sostanzialmente identico a quello del disegno *ppswr*. Per ciascuna unità i della popolazione è assegnato un numero p_i positivo che ne misura l'importanza. Anche qui, senza perdita di generalità, si assume che $p_1 + p_2 + \cdots + p_N = 1$.

Come visto nel Capitolo 3, il disegno ssr si basa sull'idea di effettuare n prove (non indipendenti), in ciascuna delle quali si seleziona con uguale probabilità una delle unità della popolazione; l'unità selezionata in una prova non può essere selezionata in nessuna delle prove successive.

Una generalizzazione immediata, che porta a disegno *ppswor*, consiste nell'effettuare n prove (non indipendenti), in ciascuna delle quali si seleziona una delle unità della popolazione. In ciascuna prova l'unità i ha probabilità di essere selezionata proporzionale a p_i. Inoltre, l'unità selezionata in una prova non può essere selezionata in nessuna delle prove successive.

Lo spazio dei campioni $\mathcal{S}$ è l'insieme di tutte le n-ple ordinate $(i_1, i_2, \ldots, i_n)$ di unità *distinte* della popolazione. In altre parole, $\mathcal{S}$ è l'insieme $\mathcal{D}_{N,n}$ delle disposizioni senza ripetizione di classe n delle unità della popolazione.

Un campione $\mathbf{s} = (i_1, i_2, \ldots, i_n)$ ha probabilità di essere selezionato pari a

$$p(\mathbf{s}) = p_{i_1} \frac{p_{i_2}}{1 - p_{i_1}} \cdots \frac{p_{i_n}}{1 - p_{i_1} - \cdots - p_{i_{n-1}}}.$$

Ovviamente, il disegno ssr corrisponde al caso speciale in cui $p_1 = \cdots = p_N = \frac{1}{N}$.

Il calcolo delle probabilità di inclusione del primo e del secondo ordine non è agevole, e non vi è una semplice formula per esprimerle. Per il caso speciale $n = 2$ si rinvia all'Esercizio 12.7.

12.4.3 Disegno di Midzuno-Lahiri

Anche il punto di partenza per la costruzione del disegno di Midzuno-Lahiri è simile a quello dei disegni *ppswr* e *ppswor*. L'assunzione di base è che per ciascuna unità i della popolazione I_N sia dato un numero p_i positivo che ne misura l'"importanza" (*size*). Senza perdita di generalità si assumerà che $p_1 + p_2 + \cdots + p_N = 1$.

Consideriamo uno schema di n "prove", in ciascuna delle quali si seleziona un'unità della popolazione, così definito.

1. Nella prima prova si seleziona un'unità della popolazione, in modo che l'unità i abbia probabilità p_i di essere selezionata.
2. Dopo aver escluso dalla popolazione l'unità selezionata nella prima prova, si seleziona dalla popolazione "residua" un campione ssr di $n-1$ unità. Formalmente, se i è selezionata nella prima prova, nelle successive $n-1$ prove si seleziona un campione ssr di $n-1$ unità da $I_N \setminus \{i\}$.

Il disegno di Midzuno-Lahiri è ottenuto mediante la riduzione dei campioni ottenuti dallo schema dianzi specificato, ovvero privando tali campioni dell'ordine di selezione delle unità (le ripetizioni sono ovviamente assenti). Lo spazio dei campioni e le probabilità dei campioni sono ricavate nell'Esercizio 12.8.

Il calcolo dele probabilità di inclusione del primo e del secondo ordine è piuttosto agevole. Per quanto riguarda quelle del primo ordine, si ha

$$\begin{aligned}
\pi_i &= Pr\left((i \textit{ selezionata } 1^\circ \textit{ prova}) \cup \begin{pmatrix} i \textit{ selezionata in una delle} \\ \textit{altre } n-1 \textit{ prove} \end{pmatrix}\right) \\
&= p_i + \sum_{\substack{j=1 \\ j \neq i}}^{N} Pr\left((j \textit{ selezionata } 1^\circ \textit{ prova}) \cap \begin{pmatrix} i \textit{ selezionata nelle} \\ \textit{altre } n-1 \textit{ prove} \end{pmatrix}\right) \\
&= p_i + \sum_{\substack{j=1 \\ j \neq i}}^{N} p_j \, Pr\begin{pmatrix} i \textit{ selezionata in un campione} \\ \textit{ssr di } n-1 \textit{ unità di } I_N \setminus \{j\}) \end{pmatrix} \\
&= p_i + \sum_{\substack{j=1 \\ j \neq i}}^{N} p_j \frac{n-1}{N-1} \\
&= \frac{N-n}{N-1} p_i + \frac{n-1}{N-1}. \qquad (12.42)
\end{aligned}$$

Per quanto riguarda invece due unità distinte i, j, esse possono essere selezionate solo in uno dei seguenti tre casi:

– nella prima prova si seleziona i, e in una delle altre $n-1$ prove si seleziona j;
– nella prima prova si seleziona j, e in una delle altre $n-1$ prove si seleziona i;
– nella prima prova si seleziona una qualsiasi unità $k \neq i,\ j$, e in due delle altre $n-1$ prove si selezionano $i,\ j$.

Ragionando come per le probabilità di inclusione del primo ordine, si può pertanto scrivere:

$$\begin{aligned}\pi_{ij} &= p_i \frac{n-1}{N-1} + p_j \frac{n-1}{N-1} + \sum_{\substack{k=1 \\ k \neq i,\, j}}^{N} p_k \frac{(n-1)(n-2)}{(N-1)(N-2)} \\ &= p_i \frac{n-1}{N-1} + p_j \frac{n-1}{N-1} + (1 - p_i - p_j) \frac{(n-1)(n-2)}{(N-1)(N-2)} \\ &= \frac{(N-n)(n-1)}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-1)(n-2)}{(N-1)(N-2)}. \end{aligned} \tag{12.43}$$

Tra probabilità di inclusione del primo e del secondo ordine sussiste un'interessante disuguaglianza. È infatti facile provare (Esercizio 12.9) che vale la relazione:

$$\pi_{ij} \leqslant \pi_i \pi_j \quad \textit{se } i \neq j. \tag{12.44}$$

12.5 Interpretazione geometrica dei disegni campionari*

I disegni campionari possiedono un'interessante interpretazione geometrica, utile anche per meglio comprendere aspetti relativi alla loro implementazione. Come già detto in precedenza, la conoscenza del campione $\mathbf{s}$ equivale alla conoscenza del vettore $\boldsymbol{\delta}(\mathbf{s})$ le cui componenti sono gli indicatori di presenza-assenza delle unità della popolazione nel campione $\mathbf{s}$. Il vettore $\boldsymbol{\delta}(\mathbf{s})$ è un vettore a N componenti ciascuna delle quali è uguale a 1 o a 0. Geometricamente, $\boldsymbol{\delta}(\mathbf{s})$ è un *vertice dell'ipercubo* $[0,\, 1]^N$ N-dimensionale di lato unitario e vertici opposti $(0,\, 0,\, \ldots,\, 0)$ e $(1,\, 1,\, \ldots,\, 1)$. Nella Fig. 12.1, tratta (con qualche variante) da Tillé (2006), è rappresentato il caso $N = 3$.

A ciascun campione $\mathbf{s} \in \mathcal{S}$ corrisponde un vertice $\boldsymbol{\delta}(\mathbf{s})$ dell'ipercubo $[0,\, 1]^N$. Pertanto, selezionare un campione $\mathbf{s}$ equivale a selezionare un "valore" del vettore $\boldsymbol{\delta}(\mathbf{s})$, che a sua volta equivale a selezionare uno dei vertici dell'ipercubo $[0,\, 1]^N$. In simboli:

Si seleziona il campione $\mathbf{s}$
se e solo se
si seleziona il vettore $\boldsymbol{\delta}(\mathbf{s})$
se e solo se
si seleziona il vertice dell'ipercubo $[0,\, 1]^N$ *corrispondente a* $\boldsymbol{\delta}(\mathbf{s})$.

Dato un disegno campionario $(\mathcal{S},\ p(\cdot))$, i campioni $\mathbf{s} \in \mathcal{S}$ di probabilità positiva corrispondono ai vertici dell'ipercubo $[0,\,1]^N$ effettivamente selezionabili. L'ampiezza effettiva del campione

$$\nu(\mathbf{s}) = \sum_{i=1}^{N} \delta(i;\ \mathbf{s})$$

è null'altro che il numero di componenti di $\boldsymbol{\delta}(\mathbf{s})$ pari a 1. In particolare, se il disegno è ad ampiezza effettiva costante n, i soli vertici di $[0,\,1]^N$ selezionabili sono quelli in cui n coordinate sono uguali a 1, e $N - n$ sono uguali a 0.

Il vettore $\boldsymbol{\pi}$ delle probabilità di inclusione del primo ordine

$$\boldsymbol{\pi} = \sum_{\mathbf{s}\in\mathcal{S}} \boldsymbol{\delta}(\mathbf{s})\, p(\mathbf{s})$$

è una *combinazione lineare convessa* dei vertici dell'ipercubo $[0,\,1]^N$ corrispondenti ai campioni $\mathbf{s} \in \mathcal{S}$. Ogni vertice ha un peso pari alla probabilità del campione a cui corrisponde.

Esempio 12.9. Si consideri una popolazione di $N = 3$ unità, $I_3 = \{1,\,2,\,3\}$, e si supponga che i campioni siano

$$\mathbf{s}_1 = \{1,\,2\},\ \mathbf{s}_2 = \{1,\,3\},\ \mathbf{s}_3 = \{2,\,3\} \tag{12.45}$$

con probabilità

$$p(\mathbf{s}_1) = 0.2,\ p(\mathbf{s}_2) = 0.3,\ p(\mathbf{s}_3) = 0.5.$$

La rappresentazione geometrica di questo disegno campionario è in Fig. 12.1.

Il vettore delle probabilità di inclusione del primo ordine è uguale a

$$\boldsymbol{\pi} = \begin{bmatrix} 0.5 \\ 0.7 \\ 0.8 \end{bmatrix}$$

ed è un punto del triangolo di vertici

$$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},\quad \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},\quad \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

□

12.6 Quanto è "casuale" un campione casuale? Entropia di disegni campionari*

La procedura di selezione del campione adottata in tutta la presente trattazione è di tipo probabilistico, nel senso che il campione è selezionato in base ad

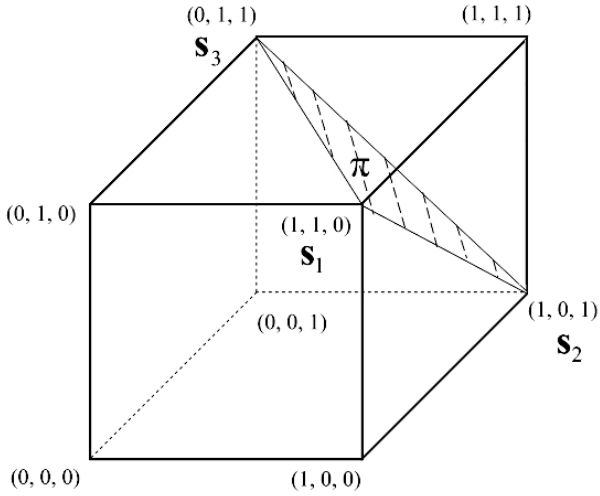


Fig. 12.1 Interpretazione geometrica di un disegno campionario

una legge di probabilità (il disegno campionario) nota *a priori* perché decisa dallo statistico. In questo senso si parla di "campione casuale". In connessione con questa nozione, sorge spontanea una domanda: "È possibile misurare il grado di casualità di un campione?"

La nozione di casualità non è propria dello specifico campione osservato, quanto piuttosto della procedura con cui esso è scelto, ossia della legge di probabilità in base alla quale è stato selezionato. Pertanto, quel che ha effettivamente senso misurare è il "grado di casualità" di un disegno campionario. Intuitivamente, un disegno campionario è tanto più "casuale" quanto più si è incerti sul campione effettivamente selezionato. Pertanto, in termini equivalenti, ma di più facile intelligibilità, l'obiettivo è quello di misurare il "grado di incertezza" derivante dalla selezione di un campione mediante un disegno campionario.

La più importante misura di incertezza è l'*entropia*, definita come

$$H = -\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \log p(\mathbf{s}) = -E[\log p(\mathbf{s})]. \tag{12.46}$$

Se $p(\mathbf{s}) = 0$, si adotterà d'ora in poi la convenzione $0 \times \infty = 0$.

È facile vedere che H assume solo valori non negativi ($H \geqslant 0$), e che $H = 0$ se e solo se vi è un unico campione $\mathbf{s}$ di probabilità 1. Un disegno di questo tipo è detto *ragionato* (*purposive*). Esso è totalmente privo di incertezza, in quanto è perfettamente noto *a priori* quale sarà il campione selezionato. Lievemente più complicato è determinare il valore massimo di H. Fissato lo spazio $\mathcal{S}$ dei campioni di unità, e detto $|\mathcal{S}|$ il numero di campioni in $\mathcal{S}$, è comunque facile provare (come conseguenza dell'Esercizio 12.14) che l'entropia H assume valore massimo se tutti i campioni in $\mathcal{S}$ hanno la stessa probabilità di essere

selezionati: $P(\mathbf{s}) = 1/|\mathcal{S}|$ per ciascun $\mathbf{s} \in \mathcal{S}$. Sul piano intuitivo, è questo il caso in cui si è massimamente incerti sul campione che sarà selezionato. In particolare, se l'insieme $\mathcal{S}$ dei campioni selezionabili è la famiglia $\mathcal{C}_{N,n}$ di tutti i sottoinsiemi di n unità distinte delle N che compongono la popolazione (famiglia delle combinazioni senza ripetizione di classe n), allora il disegno campionario di massima entropia è quello semplice senza ripetizione.

L'incertezza sul risultato della selezione di un campione mediante un disegno campionario può anche essere interpretata in termini di "ricchezza" del disegno campionario. Un disegno campionario $(\mathcal{S},\ p(\cdot))$, in altri termini, è tanto più ricco quanto più "ampio" è lo spazio dei campioni (di unità) $\mathcal{S}$, e quanto più tali campioni sono selezionabili con probabilità tra loro "vicine".

Esempio 12.10. Si consideri una popolazione finita $I_3 = \{1, 2, 3\}$ composta da $N = 3$ unità, e siano $(\mathcal{S}_1,\ p_1(\cdot))$, $(\mathcal{S}_2,\ p_2(\cdot))$ due disegni campionari definiti come:

$$\begin{aligned}
&\mathcal{S}_1 = \mathcal{S}_2 = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\};\ \ \mathbf{s}_1 = \{1, 2\},\ \mathbf{s}_2 = \{1, 3\},\ \mathbf{s}_3 = \{2, 3\}\\
&p_1(\mathbf{s}_1) = 0.8,\ p_1(\mathbf{s}_2) = 0.1,\ p_1(\mathbf{s}_3) = 0.1;\\
&p_2(\mathbf{s}_1) = 0.4,\ p_2(\mathbf{s}_2) = 0.3,\ p_2(\mathbf{s}_3) = 0.3.
\end{aligned}$$

Il disegno $(\mathcal{S}_1,\ p_1(\cdot))$ ha entropia

$$H_1 = -(0.8 \log 0.8 + 0.1 \log 0.1 + 0.1 \log 0.1) = 0.636$$

mentre il disegno $(\mathcal{S}_2,\ p_2(\cdot))$ ha entropia

$$H_2 = -(0.4 \log 0.4 + 0.3 \log 0.3 + 0.3 \log 0.3) = 1.088.$$

Il fatto che H_2 sia più grande di H_1 non è sorprendente, in quanto è intuitivamente evidente che l'incertezza relativa al campione selezionato è maggiore per il disegno $(\mathcal{S}_2,\ p_2(\cdot))$ che per $(\mathcal{S}_1,\ p_1(\cdot))$. □

A parità di caratteristiche, quali l'insieme $\mathcal{S}$ dei campioni selezionabili, le probabilità di inclusione del primo e del secondo ordine, la facilità di implementazione mediante algoritmi numericamente efficienti, etc., è in generale preferibile usare un disegno campionario con elevata entropia. Per fornire un argomento intuitivo a supporto di quest'affermazione, consideriamo il seguente esempio.

Esempio 12.11. Si consideri una popolazione finita $I_4 = \{1, 2, 3, 4\}$ di $N = 4$ unità, di cui indichiamo con $y_1,\ y_2,\ y_3,\ y_4$ le modalità etichettate. Supponiamo di dover stimare la media $\mu_y = (y_1 + y_2 + y_3 + y_4)/4$ della popolazione. In assenza di informazioni *a priori*, una scelta molto naturale consiste nel selezionare un campione mediante un disegno (semplice senza ripetizione di numerosità $n = 2$) $(\mathcal{S}_1,\ p_1(\cdot))$, definito da:

$$\begin{aligned}
&\mathcal{S}_1 = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\};\\
&p_1(\{1, 2\}) = p_1(\{1, 3\}) = \cdots = p_1(\{3, 4\}) = \tfrac{1}{6}.
\end{aligned}$$

Tutte le unità hanno probabilità di inclusione del primo ordine pari a 1/2.

In questo caso la media campionaria $\overline{y}_{\mathbf{s}} = \sum_{i \in \mathbf{s}} y_i/2$ è uno stimatore corretto della media della popolazione. Per esaminare più da vicino il comportamento dello stimatore $\overline{y}_{\mathbf{s}}$, consideriamo la Tabella 12.1, in cui sono riportati i valori di $\overline{y}_{\mathbf{s}}$ corrispondenti ai diversi campioni, e le relative probabilità.

Tabella 12.1 Costruzione dello stimatore $\overline{y}_{\mathbf{s}}$ per il disegno $(\mathcal{S}_1,\ p_1(\cdot))$

Campione	*Probabilità*	$\overline{y}_{\mathbf{s}}$
{1, 2}	1/6	$\frac{y_1+y_2}{2}$
{1, 3}	1/6	$\frac{y_1+y_3}{2}$
{1, 4}	1/6	$\frac{y_1+y_4}{2}$
{2, 3}	1/6	$\frac{y_2+y_3}{2}$
{2, 4}	1/6	$\frac{y_2+y_4}{2}$
{3, 4}	1/6	$\frac{y_3+y_4}{2}$

Supponiamo ora di disporre *a priori* di un'informazione aggiuntiva, ovvero che le unità 1, 2 hanno valori y molto simili tra loro, e che le unità 3, 4 hanno anch'esse valori y molto simili, anche se assai diversi dai precedenti. Per semplificare un po', si può ammettere che y_1 e y_2 siano uguali e "piccoli", mentre y_3 e y_4 siano uguali e "grandi". Mantenendo una numerosità campionaria $n = 2$, conviene in questo caso limitarsi ai soli campioni che contengono una tra le unità 1, 2 e una tra le unità 3, 4. Un disegno di campionamento ragionevole per questa situazione è quello di seguito riportato, indicato con $(\mathcal{S}_2,\ p_2(\cdot))$.

$$\mathcal{S}_2 = \{\{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}\};$$
$$p_2(\{1, 3\}) = p_2(\{1, 4\}) = p_2(\{2, 3\} = p_2(\{2, 4\}) = \tfrac{1}{4}.$$

Anche questo disegno ha ampiezza effettiva costante $n = 2$, e dà a tutte le unità della popolazione probabilità di inclusione 1/2. Anche in questo caso la media campionaria è uno stimatore corretto della media della popolazione. Il suo comportamento è riassunto nella Tabella 12.2.

È facile verificare che il primo disegno ha entropia $H_1 = 1.79$, mentre il secondo disegno ha entropia $H_2 = 1.49$.

Se l'informazione *a priori* $y_1 = y_2$ "piccoli", $y_3 = y_4$ "grandi" è corretta, e se si stima la media della popolazione con la media campionaria, il disegno $(\mathcal{S}_2,\ p_2(\cdot))$ fornisce risultati migliori di $(\mathcal{S}_1,\ p_1(\cdot))$, in quanto, come indicato nella successiva Tabella 12.3, fornisce *sempre* il "vero valore" di μ_y.

La maggior efficienza di stima, in questo caso, è quindi ottenuta usando un disegno campionario $(\mathcal{S}_2,\ p_2(\cdot))$ di entropia più piccola di quella del disegno $(\mathcal{S}_1,\ p_1(\cdot))$.

Tabella 12.2 Costruzione dello stimatore $\overline{y}_{\mathbf{s}}$ per il disegno $(\mathcal{S}_2,\ p_2(\cdot))$

Campione	*Probabilità*	$\overline{y}_{\mathbf{s}}$
{1, 3}	1/4	$\frac{y_1+y_3}{2}$
{1, 4}	1/4	$\frac{y_1+y_4}{2}$
{2, 3}	1/4	$\frac{y_2+y_3}{2}$
{2, 4}	1/4	$\frac{y_2+y_4}{2}$

Tabella 12.3 Caratteristiche dello stimatore $\overline{y}_{\mathbf{s}}$ per i disegni $(\mathcal{S}_1,\ p_1(\cdot))$, $(\mathcal{S}_2,\ p_2(\cdot))$ nel caso $y_1 = y_2$ "piccoli", $y_3 = y_4$ "grandi"

Disegno $(\mathcal{S}_1,\ p_1(\cdot))$			*Disegno* $(\mathcal{S}_2,\ p_2(\cdot))$		
Campione	*Probabilità*	*Stima* $\overline{y}_{\mathbf{s}}$	*Campione*	*Probabilità*	*Stima* $\overline{y}_{\mathbf{s}}$
{1, 2}	1/6	*Sottostima*			
{1, 3}	1/6	*OK*	{1, 3}	1/4	*OK*
{1, 4}	1/6	*OK*	{1, 4}	1/4	*OK*
{2, 3}	1/6	*OK*	{2, 3}	1/4	*OK*
{2, 4}	1/6	*OK*	{2, 4}	1/4	*OK*
{3, 4}	1/6	*Sovrastima*			

Supponiamo ora che l'informazione *a priori* $y_1 = y_2$ "piccoli", $y_3 = y_4$ "grandi", sulla base della quale si è costruito il disegno $(\mathcal{S}_2,\ p_2(\cdot))$, sia scorretta. Cosa succede in questo caso? Per essere concreti, consideriamo ancora i due disegni di campionamento $(\mathcal{S}_1,\ p_1(\cdot))$, $(\mathcal{S}_2,\ p_2(\cdot))$, ma supponiamo adesso che $y_1 = y_3$ siano "piccoli", e che $y_2 = y_4$ siano "grandi". Le caratteristiche della media campionaria $\overline{y}_{\mathbf{s}}$ sono riassunte nella successiva Tabella 12.4.

Tabella 12.4 Caratteristiche dello stimatore $\overline{y}_{\mathbf{s}}$ per i disegni $(\mathcal{S}_1,\ p_1(\cdot))$, $(\mathcal{S}_2,\ p_2(\cdot))$ nel caso $y_1 = y_3$ "piccoli", $y_2 = y_4$ "grandi"

Disegno $(\mathcal{S}_1,\ p_1(\cdot))$			*Disegno* $(\mathcal{S}_2,\ p_2(\cdot))$		
Campione	*Probabilità*	*Stima* $\overline{y}_{\mathbf{s}}$	*Campione*	*Probabilità*	*Stima* $\overline{y}_{\mathbf{s}}$
{1, 2}	1/6	*OK*			
{1, 3}	1/6	*Sottostima*	{1, 3}	1/4	*Sottostima*
{1, 4}	1/6	*OK*	{1, 4}	1/4	*OK*
{2, 3}	1/6	*OK*	{2, 3}	1/4	*OK*
{2, 4}	1/6	*Sovrastima*	{2, 4}	1/4	*Sovrastima*
{3, 4}	1/6	*OK*			

Se si usa il disegno $(\mathcal{S}_1,\ p_1(\cdot))$ si ha con probabilità $2/3$ una stima esattamente uguale alla media della popolazione, e con probabilità $1/3$ si commette un (serio) errore di stima (sottostima o sovrastima). Se invece si usa il disegno $(\mathcal{S}_2,\ p_2(\cdot))$ si ha con probabilità $1/2$ una stima esattamente uguale alla media della popolazione, e con probabilità $1/2$ si commette un (serio) errore di stima (sottostima o sovrastima). Quindi, se da un lato il disegno $(\mathcal{S}_2,\ p_2(\cdot))$ fornisce risultati migliori di $(\mathcal{S}_1,\ p_1(\cdot))$ quando è corretta l'informazione *a priori* usata per la sua costruzione, dall'altro fornisce risultati peggiori quando tale informazione non trova riscontro nella realtà. La ragione per cui ciò accade è che il disegno $(\mathcal{S}_1,\ p_1(\cdot))$ è più "ricco" di $(\mathcal{S}_2,\ p_2(\cdot))$, in quanto il relativo spazio dei campioni comprende anche i campioni $\{1, 2\}$, $\{3, 4\}$, e con probabilità di selezione non trascurabili. La presenza di tali campioni, e la corrispondente maggior ricchezza dello spazio dei campioni (di unità), in un certo senso, fa in modo che le caratteristiche dello stimatore $\overline{y}_{\mathbf{s}}$ non siano troppo negativamente influenzate da casi in cui le informazioni a priori sulla popolazione sono scorrette, e molto lontane dalla realtà. La maggior entropia del disegno $(\mathcal{S}_1,\ p_1(\cdot))$ è quindi una forma di "protezione" delle proprietà dello stimatore $\overline{y}_{\mathbf{s}}$ rispetto a popolazioni "estreme" (molto divergenti dalle informazioni *a priori* di cui si dispone), garantendo allo stimatore stesso una certa *robustezza*. D'altra parte tale robustezza è ottenuta al prezzo di una minore efficienza nei casi in cui le informazioni *a priori* sulla popolazione risultano corrette. □

Quel che emerge dall'Esempio 12.11 porta ad alcune considerazioni di carattere generale. Supponiamo di disporre di informazioni *a priori* sulla popolazione di interesse. Esattamente come nell'Es. 12.11, tali informazioni possono essere utilizzate per la costruzione di un disegno campionario. In genere, tale disegno è caratterizzato da una bassa entropia, in quanto dà probabilità di selezione alta ai campioni coerenti con l'informazione stessa, e probabilità di selezione molto piccola ai restanti campioni. Se l'informazione *a priori* utilizzata si rivela corretta, si avrà un'alta efficienza di stima. D'altra parte, nel caso in cui essa si riveli scorretta, si è esposti a gravi errori di stima. La maggior efficienza è quindi ottenuta a prezzo di una scarsa robustezza.

Un disegno campionario ad alta entropia è invece un disegno "ricco di campioni", nel senso che vi saranno "molti" campioni con probabilità di selezione "vicine" tra loro e abbastanza alte, e "pochi" campioni con probabilità di essere selezionati "piccola" (compatibilmente con eventuali vincoli riguardanti grandezze quali le probabilità di inclusione del primo e/o del secondo ordine, o altro ancora). A causa di questa sua caratteristica, un disegno ad alta entropia garantisce una forma di *robustezza* di stima nel caso di popolazione "lontana" dalle informazioni *a priori* di cui si dispone. Da questo punto di vista, un'alta entropia del disegno campionario è una caratteristica positiva, desiderabile. Vi è ovviamente il rovescio della medaglia: la maggior robustezza è ottenuta al prezzo di uno sfruttamento meno efficiente delle informazioni *a priori* sulla popolazione di interesse.

12.7 Calcolo approssimato delle probabilità di inclusione del secondo ordine

In molti disegni campionari di uso concreto, specialmente in quelli che verranno esposti nel Capitolo 15, il calcolo delle probabilità di inclusione può essere estremamente complicato. In linea di principio, a meno di casi speciali, il calcolo delle probabilità di inclusione richiede l'enumerazione dei campioni dello spazio $\mathcal{S}$, e questo è quasi sempre troppo oneroso sul piano computazionale. In linea di principio, questa difficoltà sussiste per il calcolo delle probabilità di inclusione sia del primo che del secondo ordine. Tuttavia, per ragioni che saranno chiare nel Capitolo 15, questo problema riguarda soprattutto il calcolo di quelle del secondo ordine. Per questa ragione è di notevole interesse cercare una qualche forma approssimata per le π_{ij}.

La più semplice tra le varie approssimazioni per π_{ij} proposte in letteratura è la seguente (introdotta in cfr. Hájek (1981) per alcuni speciali disegni campionari)

$$\pi_{ij} \approx \pi_i \, \pi_j \left(1 - \frac{(1-\pi_i)\,(1-\pi_j)}{d} \right) \tag{12.47}$$

dove si è posto

$$d = \sum_{i=1}^{N} \pi_i \,(1-\pi_i).$$

L'ipotesi di base è che il valore di d sia "grande".

Il pregio maggiore della (12.47) è indubbiamente la sua semplicità. La sua accuratezza, però, non è sempre delle migliori. In generale, l'approssimazione (12.47) fornisce risultati tanto migliori quanto più elevata è l'entropia del disegno campionario: più alta è l'entropia, migliore è la qualità dell'approssimazione.

Esempio 12.12. Se il disegno campionario è semplice senza ripetizione si ha $\pi_i = \frac{n}{N}$, $\pi_{ij} = \frac{n\,(n-1)}{N\,(N-1)}$, e

$$d = \sum_{i=1}^{N} \frac{n}{N} \left(1 - \frac{n}{N} \right) = n \left(1 - \frac{n}{N} \right).$$

Si può quindi scrivere

$$\begin{aligned} \pi_i \, \pi_j \left(1 - \frac{(1-\pi_i)\,(1-\pi_j)}{d} \right) &= \frac{n^2}{N^2} \left(1 - \frac{1 - \frac{n}{N}}{n} \right) = \frac{n}{N} \left\{ \frac{n}{N} \left(1 - \frac{1}{n} + \frac{1}{N} \right) \right\} \\ &= \frac{n}{N} \left\{ \frac{n-1}{N-1} + \frac{1}{N} \left(\frac{n}{N} - \frac{n-1}{N-1} \right) \right\} \\ &= \frac{n\,(n-1)}{N\,(N-1)} + \frac{1}{N} \, \frac{n}{N} \, \frac{N-n}{N\,(N-1)}. \end{aligned}$$

La quantità

$$err = \frac{1}{N}\,\frac{n}{N}\,\frac{N-n}{N\,(N-1)}$$

è l'*errore di approssimazione*. è facile vedere che valgono le due disuguaglianze

$$0 < err < \frac{1}{N^2}$$

per cui in questo caso l'approssimazione (12.47) è accurata. □

Una delle fonti di inaccuratezza dell'approssimazione (12.47) è che in generale non è detto che soddisfi la relazione, valida per disegni ad ampiezza effettiva costante n,

$$\sum_{\substack{j=1\\ j\neq i}}^{N} \pi_{ij} = (n-1)\pi_i \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{12.48}$$

Essendo

$$\pi_{ij} = E[\delta(i;\,\mathbf{s})\,\delta(i;\,\mathbf{s})] = C(\delta(i;\,\mathbf{s}),\,\delta(i;\,\mathbf{s})) + \pi_i\,\pi_j$$

cercare un'approssimazione per π_{ij} equivale ovviamente a cercare un'approssimazione per la covarianza

$$\Delta_{ij} = C(\delta(i;\,\mathbf{s}),\,\delta(i;\,\mathbf{s})) \tag{12.49}$$

la quale, se il disegno è ad ampiezza effettiva costante n, deve soddisfare la relazione (equivalente alla (12.48))

$$\sum_{\substack{j=1\\ j\neq i}}^{N} \Delta_{ij} = -\pi_i(1-\pi_i) \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{12.50}$$

Un'idea semplice ma efficace (cfr. Hájek (1981), pp. 24–26) consiste nell'approssimare Δ_{ij} con un prodotto:

$$\Delta_{ij} \approx \Delta^a_{ij} = -c_i\,c_j \tag{12.51}$$

dove $c_1, \ldots, c_N$ sono numeri positivi che soddisfano le (12.50). D'ora in avanti, porremo

$$C = \sum_{i=1}^{N} c_i; \quad d = \sum_{i=1}^{N} \pi_i(1-\pi_i)$$

e supporremo (per ragioni che saranno chiare tra poco) che sia $d \geqslant 3/4$.

Per soddisfare le (12.50), i numeri c_i devono essere tali che

$$-\pi_i(1-\pi_i) = c_i \sum_{\substack{j=1\\ j\neq i}}^{N} c_j = c_i(C - c_i)$$

ossia

$$c_i^2 - C\,c_i + \pi_i(1-\pi_i) = 0; \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{12.52}$$

Le (12.52) sono equazioni di secondo grado, ciascuna delle quali ha le due soluzioni

$$c_i = \frac{C}{2} \pm \sqrt{\frac{C^2}{4} - \pi_i(1-\pi_i)}.$$

La soluzione $C/2 + \sqrt{C^2/4 - \pi_i(1-\pi_i)}$ è però impossibile (Esercizio 12.16), per cui i numeri c_i sono tali che

$$c_i = \frac{C}{2} - \sqrt{\frac{C^2}{4} - \pi_i(1-\pi_i)} \;\; \textit{per ciascuna unità } i = 1, \ldots, N. \tag{12.53}$$

Per determinare completamente i numeri c_i bisogna calcolare C. Sommando le (12.53) rispetto a i, si ha

$$C = \frac{CN}{2} - \sum_{i=1}^{N} \sqrt{\frac{C^2}{4} - \pi_i(1-\pi_i)}$$

che equivale a

$$C - \frac{2}{N-2} \sum_{i=1}^{N} \sqrt{\frac{C^2}{4} - \pi_i(1-\pi_i)} = 0. \tag{12.54}$$

L'equazione (12.54) non può essere risolta esplicitamente. è necessario invece ricorrere ad un metodo numerico. Il più semplice è il *metodo delle bisezioni*, di seguito descritto.

Per l'implementazione del metodo delle bisezioni è necessario partire da due numeri C_s, C_d (facilmente determinabili per tentativi) tali che

$$C_s - \frac{2}{N-2} \sum_{i=1}^{N} \sqrt{\frac{C_s^2}{4} - \pi_i(1-\pi_i)} < 0, \quad C_d - \frac{2}{N-2} \sum_{i=1}^{N} \sqrt{\frac{C_d^2}{4} - \pi_i(1-\pi_i)} > 0.$$

- *Passo* 0. Inizializzazione. Porre $s_x = C_s$, $d_x = C_d$, e fissare una soglia $\delta > 0$ ''piccola''. Andare al Passo 1.
- *Passo* 1. Se $|d_x - s_x| < \delta$ andare al Passo 3. Altrimenti, porre $C = (s_x + d_x)/2$ e andare al Passo 2.
- *Passo* 2. Calcolare

  $$C - \frac{2}{N-2} \sum_{i=1}^{N} \sqrt{\frac{C^2}{4} - \pi_i(1-\pi_i)}.$$

  Se tale quantità è maggiore di 0, porre $d_x = C$ e andare al Passo 1. Se invece è minore di 0, porre $s_x = C$ e andare al Passo 1.
- *Passo* 3. Arresto. Prendere il valore $C = (s_x + d_x)/2$ come soluzione dell'equazione (12.54).

Qualunque sia l'algoritmo numerico (bisezioni o altro) usato per risolvere l'equazione (12.54), esso permette in ogni caso di determinare i numeri $c_1, \ldots, c_N$ i quali, tramite la (12.51), consentono di approssimare le probabiltà di inclusione π_{ij} con

$$\pi_{ij} \approx \pi_{ij}^a = \pi_i \, \pi_j - c_i \, c_j. \tag{12.55}$$

Si può anche dimostrare (cfr. Hájek (1981)) che per d "grande" l'approssimazione (12.55) si riduce alla forma semplificata (12.47). Inoltre, anche in questo caso la qualità dell'approssimazione è legata all'entropia del disegno campionario, nel senso che per disegni campionari a bassa entropia l'approssimazione (12.55) darà risultati cattivi, mentre per disegni ad alta entropia fornirà risultati buoni.

Esempio 12.13. Se il disegno campionario è semplice senza ripetizione si ha, come già visto, $\pi_i = \frac{n}{N}$ e $\pi_{ij} = \frac{n\,(n-1)}{N\,(N-1)}$. In questo caso è evidente che $c_1 = \cdots = c_N = c$, per cui è anche $C = N\,c$. L'equazione (12.52) assume la forma

$$c^2(N-1) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

da cui si desume che

$$c_i \, c_j = c^2 = \frac{1}{N-1}\,\frac{n}{N}\left(1 - \frac{n}{N}\right).$$

Si ha pertanto:

$$\pi_{ij}^a = \pi_i \, \pi_j - c_i \, c_j = \left(\frac{n}{N}\right)^2 - \frac{1}{N-1}\,\frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n\,(n-1)}{N\,(N-1)}$$

ossia l'approssimazione (12.55) fornisce il valore esatto delle probabilità di inclusione del secondo ordine. □

Per altri esempi che coinvolgono la (12.55) si rinvia a Hájek (1981) (pp. 90–91) e Bondesson *e altri* (2006).

12.8 Implementazione di disegni campionari: aspetti generali

Un aspetto molto importante legato all'uso dei disegni campionari è la loro *implementazione*, ovvero l'effettiva selezione di un campione in base a quel disegno. Affinché un disegno campionario possa realmente essere utilizzato è necessario disporre di un qualche *schema*, di un qualche *algoritmo* numericamente efficiente per la sua implementazione. Come visto nel Capitolo 2, il metodo in linea di principio più semplice per implementare un disegno campionario è quello dell'inversione della funzione di ripartizione. Purtroppo, come già scritto ed esemplificato, esso non è quasi mai numericamente efficiente, in quanto si basa sull'enumerazione dei campioni, operazione in genere talmente lunga e onerosa da essere praticamente irrealizzabile.

Qui di seguito vengono brevemente esaminate alcune tipologie di algoritmi (schemi) per l'implementazione di disegni campionari. Per semplicità ci si limiterà esclusivamente a *disegni non ordinati e senza ripetizioni*. Per una trattazione completa, e decisamente brillante, si rinvia al volume di Tillé (2006).

12.8.1 Schemi basati su estrazioni successive

Questo tipo di schemi si applica esclusivamente a disegni ad ampiezza costante n, e si basa su n "prove" in ciascuna delle quali si "estrae" un'unità della popolazione. Se $(\mathcal{S}, p(\cdot))$ è un disegno ad ampiezza effettiva costante n lo schema più semplice di estrazioni successive è di seguito descritto. Indicheremo con $\boldsymbol{d}$

$$\boldsymbol{d} = \begin{bmatrix} d_1 \\ d_2 \\ \cdots \\ d_N \end{bmatrix} \tag{12.56}$$

un vettore a N componenti in cui ciascun elemento d_i è uguale o a 0, o a 1. Inoltre, indicheremo con $\mathbf{0}$ un vettore a N componenti tutte uguali a 0.

- *Passo* 1. Inizializzazione. Porre $\boldsymbol{d} = \mathbf{0}$, $t = 1$. Andare al Passo 2.
- *Passo* 2. Se $t > n$ andare al Passo 4. Altrimenti, per ciascuna unità i della popolazione, calcolare

$$p_i = \begin{cases} 0 \text{ se } d_i = 1 \\ \frac{1}{n-t+1} Pr\left(\delta(i;\, \mathbf{s}) = 1 \,|\delta(k;\, \mathbf{s}) = 1 \begin{array}{l} \text{per tutte le unità } k \\ \text{tali che } d_k \end{array} = 1\right) \end{cases}$$

  e porre $P_0 = 0$, $P_i = p_1 + p_2 + \cdots + p_i$ per $i = 1, 2, \ldots, N$. Andare al Passo 3.
- *Passo* 3. Generare una variabile aleatoria U con distribuzione uniforme in $[0, 1]$. Se $P_{i-1} \leqslant U < P_i$ selezionare l'unità i e porre $d_i = 1$. Incrementare t di 1. Andare al Passo 2.
- *Passo* 4. Arresto. Il campione $\mathbf{s}$ è formato dalle n unità i tali che $d_i = 1$.

Ad esempio, nel disegno semplice senza ripetizione (ssr) è facile verificare che p_i è uguale a $(n - t + 1)/(N - t + 1)$ se $d_i = 0$, così che l'algoritmo sopra descritto è semplicissimo da implementare.

Lo schema basato su estrazioni successive non è necessariamente numericamente efficiente, in quanto il calcolo delle probabilità p_i può richiedere l'enumerazione dei campioni dell'insieme $\mathcal{S}$.

Lo schema di estrazioni successive può essere facilmente esteso al caso di disegni ordinati e/o con ripetizioni, ferma restando un'ampiezza campionaria costante. Sono ad esempio di questo tipo gli schemi brevemente delineati per implementare i disegni *ppswor*, *ppswr*, e di Midzuno-Lahiri.

12.8.2 Schemi basati su algoritmi sequenziali

L'idea di base degli schemi di tipo sequenziale è quella di effettuare una "prova" per ciascuna unità della popolazione. Il risultato di tale prova può essere

o l'inclusione dell'unità nel campione, o la sua non inclusione. Al solito, per ragioni di semplicità ci limiteremo esclusivamente a disegni campionari non ordinati e senza ripetizioni.

Dato il disegno campionario $(\mathcal{S}, p(\cdots))$, consideriamo le N variabili indicatrici $\delta(1; \mathbf{s})$, ..., $\delta(N; \mathbf{s})$. Consideriamo inoltre N quantità d_1, ..., d_N, ognuna delle quali può essere uguale a 0 oppure a 1. In simboli:

$$d_1 = 0, 1; \;\; d_2 = 0, 1, \ldots, \;\; d_N = 0, 1.$$

Infine, per ogni unità i $(= 1, \ldots, N)$ e per ogni possibile scelta dei valori di $d_1, \ldots, d_N$ poniamo

$$\begin{aligned} S_i(d_1, \ldots, d_i) &= \textit{Insieme dei campioni } \mathbf{s} \in \mathcal{S} \textit{ tali che } \delta(1; \mathbf{s}) \\ &= d_1, \ldots \delta(i; \mathbf{s}) = d_i. \end{aligned}$$

Ad es., $S_1(1)$ è l'insieme dei campioni $\mathbf{s}$ che contengono l'unità 1, $S_1(0)$ è l'insieme dei campioni $\mathbf{s}$ che non contengono l'unità 1, $S_3(1, 0, 1)$ è l'insieme dei campioni $\mathbf{s}$ che contengono l'unità 1, non contengono l'unità 2, e contengono l'unità 3.

Poniamo infine

$$q_1(d_1) = Pr(\delta(1; \mathbf{s}) = d_1) = \begin{cases} \pi_1 \textit{ se } d_1 = 1 \\ 1 - \pi_1 \textit{ se } d_1 = 0 \end{cases}$$

e in generale, per ciascuna delle altre unità $i = 2, \ldots, N$,

$$\begin{aligned} q_i(d_i) &= Pr(\delta(i; \mathbf{s}) = d_i \,|\delta(1; \mathbf{s}) = d_1, \ldots, \delta(i-1; \mathbf{s}) = d_{i-1}) \\ &= \begin{cases} Pr(\delta(i; \mathbf{s}) = 1 \,|\delta(1; \mathbf{s}) = d_1, \ldots, \delta(i-1; \mathbf{s}) = d_{i-1}) \textit{ se } d_i = 1 \\ Pr(\delta(i; \mathbf{s}) = 0 \,|\delta(1; \mathbf{s}) = d_1, \ldots, \delta(i-1; \mathbf{s}) = d_{i-1}) \textit{ se } d_i = 0\,. \end{cases} \end{aligned}$$

Naturalmente, è $q_i(0) = 1 - q_i(1)$. La probabilità condizionata $q_i(1)$ può essere interpretata come probabilità di inclusione dell'untà i *condizionata* ai valori di $\delta(1\,, \mathbf{s})$, ..., $\delta(i-1\,, \mathbf{s})$, ossia all'aver incluso nel campione le unità dalla 1 alla $i-1$ tali che $d = 1$, e al non aver incluso nel campione le unità dalla 1 alla $i-1$ tali che $d = 0$. In termini espliciti, $q_i(d_i)$ è pari a

$$q_i(d_i) = \frac{\sum_{\mathbf{s} \in S_i(d_1, \ldots, d_i)} p(\mathbf{s})}{\sum_{\mathbf{s} \in S_{i-1}(d_1, \ldots, d_{i-1})} p(\mathbf{s})}.$$

L'algoritmo sequenziale di selezione del campione, descritto qui di seguito, si basa su un principio semplicissimo. Si "mettono in fila" le unità, dalla 1 alla N. L'unità 1 è selezionata con probabilità pari alla propria probabilità di inclusione, l'unità 2 è selezionata con probabilità pari alla propria probabilità di inclusione condizionata all'aver incluso o meno l'unità 1 nel campione, e così via.

– *Passo* 1. `Inizializzazione. Porre` $i = 1$, $d_1 = 0$, ..., $d_N = 0$. `Andare al Passo 2.`

- *Passo* 2. Se $i > N$ andare al Passo 4. Altrimenti, calcolare $q_i(1)$, $q_i(0)$, e andare al Passo 3.
- *Passo* 3. Generare una variabile aleatoria U con distribuzione uniforme in $[0, 1]$. Se $U \leqslant q_i(1)$, porre $d_i = 1$; altrimenti porre $d_i = 0$. Incrementare i di 1. Andare al Passo 2.
- *Passo* 4. Arresto. Il campione $\mathbf{s}$ selezionato è formato dalle unità i tali che $d_i = 1$.

La validità dell'algoritmo sequenziale è semplicissima da provare (Esercizio 12.10), basta tener conto della relazione

$$p(\mathbf{s}) = q_1(d_1)\, q_2(d_2) \cdots q_N(d_N). \tag{12.57}$$

L'efficienza computazionale dell'algoritmo di tipo sequenziale dipende dalla facilità o difficoltà di calcolo delle probabilià condizionate $q_i(d_i)$. Se tale calcolo richiede l'enumerazione dei campioni, l'efficienza computazionale è ovviamente bassa.

A titolo di esempio osserviamo infine che nel caso del disegno semplice con ripetizione si ha

$$q_i(d_i) = \begin{cases} \frac{n - \sum_{j=1}^{i-1} d_j}{N-i+1} & \text{se } d_i = 1 \\ 1 - \frac{n - \sum_{j=1}^{i-1} d_j}{N-i+1} & \text{se } d_i = 0 \end{cases}, \quad i = 1, 2, \ldots, N.$$

12.8.3 Schemi basati su algoritmi accettazione/rifiuto

Gli schemi di tipo accettazione/rifiuto, spesso usati per implementare disegni campionari, si basano su un'idea molto semplice: selezionare in maniera "indiretta" un campione da un disegno $(\mathcal{S}_1, p_1(\cdot))$, nel seguente modo.

- Si genera un campione $\mathbf{s}$ da un dato disegno campionario $(\mathcal{S}_2, p_2(\cdot))$, da cui è "facile" selezionare campioni.
- Se il campione $\mathbf{s}$ soddisfa un'opportuna condizione, allora viene "accettato" come campione selezionato dal disegno $(\mathcal{S}_1, p_1(\cdot))$. Naturalmente, la condizione che il campione $\mathbf{s}$ deve soddisfare deve essere tale da garantire che i campioni accettati possano essere considerati come selezionati da $(\mathcal{S}_2, p_2(\cdot))$.

Nel seguito vengono brevemente esposti due dei principali algoritmi di accettazione/rifiuto. Il primo di essi è l'algoritmo di *accettazione condizionata*. Supponiamo di voler selezionare un campione dal disegno $(\mathcal{S}_1, p_1(\cdot))$, e supponiamo che il disegno $(\mathcal{S}_2, p_2(\cdot))$ sia tale che $\mathcal{S}_1 \subseteq \mathcal{S}_2$, e che

$$p_2(\mathbf{s} \,|\mathbf{s} \in \mathcal{S}_1) = p_2(\mathbf{s}) \quad \textit{per ciascun campione } \mathbf{s} \in \mathcal{S}_1. \tag{12.58}$$

Chiaramente, la probabilità condizionata $p_2(\mathbf{s} \,|\mathbf{s} \in \mathcal{S}_1)$ che appare in (12.58) si scrive come

$$p_2(\mathbf{s} \,|\mathbf{s} \in \mathcal{S}_1) = \frac{p_2(\mathbf{s})}{\sum_{\mathbf{s}' \in \mathcal{S}_1} p_2(\mathbf{s}')}.$$

L'algoritmo di accettazione condizionata, di seguito descritto, è molto semplice e intuitivo.

- *Passo* 1. Inizializzazione. Generare un campione s da $(\mathcal{S}_2, p_2(\cdot))$. Andare al Passo 2.
- *Passo* 2. Se $\mathbf{s} \in \mathcal{S}_1$, andare al Passo 1. Altrimenti, andare al Passo 3.
- *Passo* 3. Arresto. Accettare il campione s come selezionato dal disegno $(\mathcal{S}_1, p_1(\cdot))$.

L'algoritmo di accettazione condizionata seleziona ovviamente campioni dalla distribuzione condizionata $p_2(\mathbf{s} \,|\mathbf{s} \in \mathcal{S}_1)$. La sua validità poggia sulla relazione (12.58).

Si considerino, a titolo di esempio, i disegni campionari semplici senza e con ripetizioni (rispettivamente ssr e scr), ambedue di numerosità n. Chiaramente, il disegno ssr è null'altro che la riduzione del disegno $(\mathcal{S}_1, p_1(\cdot))$, in cui $\mathcal{S}_1$ è l'insieme delle disposizioni con ripetizione (n-ple ordinate) di n unità della popolazione, e ogni disposizione ha probabilità $1/\{N(N-1)\cdots(N-n+1)\}$. Per selezionare un campione mediante disegno ssr è sufficiente selezionare un campione da $(\mathcal{S}_1, p_1(\cdot))$, e privare tale campione dell'ordine. Il disegno scr svolge qui il ruolo di $(\mathcal{S}_2, p_2(\cdot))$. Chiaramente, si ha $\mathcal{S}_1 \subset \mathcal{S}_2$. È facile verificare (Esercizio 12.11) che vale la relazione (12.58), per cui per generare un campione ssr basta generare un campione scr e verificare che le unità che lo compongono siano tutte differenti. Come estensione immediata, si può verificare che per selezionare un campione con disegno *ppswor* è sufficiente generare un campione con disegno *ppswr*, e verificare che le unità che lo compongono siano tutte distinte.

L'algoritmo di accettazione condizionata è tanto più efficiente quanto più alta è la probabilità che mediante $(\mathcal{S}_2, p_2(\cdot))$ si generi un campione appartenente a $\mathcal{S}_1$. Tale probabilità definisce il *tasso di accettazione* (*acceptance rate*) dell'algoritmo, ed è pari a:

$$AR = \sum_{\mathbf{s} \in \mathcal{S}_1} p_2(\mathbf{s}). \tag{12.59}$$

Nell'esempio del disegno ssr sopra considerato è immediato constatare che

$$AR = \frac{N(N-1)\cdots(N-n+1)}{N^n} = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n-1}{N}\right).$$

Un altro algoritmo di tipo accettazione/rifiuto spesso usato è l'algoritmo di *rigetto*. Supponiamo di voler selezionare un campione dal disegno $(\mathcal{S}, p_1(\cdot))$, e supponiamo di disporre di un disegno $(\mathcal{S}, p_2(\cdot))$ tale che vi è una costante B nota per cui si ha

$$p_1(\mathbf{s}) \leqslant p_2(\mathbf{s}) \quad \textit{per ciascun campione } \mathbf{s} \in \mathcal{S}. \tag{12.60}$$

L'algoritmo di rigetto è descritto qui di seguito.

- *Passo* 1. Inizializzazione. Generare un campione $\mathbf{s}$ da $(\mathcal{S}, p_2(\cdot))$. Andare al Passo 2.
- *Passo* 2. Generare una variabile aleatoria U con distribuzione uniforme in $[0, 1]$. Se $U \leqslant \frac{p_1(\mathbf{s})}{B\, p_2(\mathbf{s})}$ andare al Passo 3. Altrimenti, andare al Passo 1.
- *Passo* 3. Arresto. Accettare il campione $\mathbf{s}$ come selezionato dal disegno $(\mathcal{S}, p_1(\cdot))$.

Ovviamente, è $B > 1$. Il numero di volte in cui è necessario generare campioni al Passo 1 per arrivare ad un'accettazione al Passo 2 è il numero di *cicli* dell'algoritmo. La validità del metodo di rigetto è provata qui di seguito.

La probabilità di accettare un dato campione $\mathbf{s}$ al Passo 1 è pari a:

$$\begin{aligned} Pr(Accettazione | Il\ campione\ generato\ è\ \mathbf{s}) &= Pr\left(U \leqslant \frac{p_1(\mathbf{s})}{B\, p_2(\mathbf{s})}\right) \\ &= \frac{p_1(\mathbf{s})}{B\, p_2(\mathbf{s})} \end{aligned} \tag{12.61}$$

per cui la probabilità di generare un dato campione $\mathbf{s}$ e immediatamente accettarlo in un ciclo è uguale a

$$\begin{aligned} &Pr(Generare\ il\ campione\ \mathbf{s}\ immediatamente\ accettato) \\ &= Pr(Generare\ il\ campione\ \mathbf{s})\, Pr(Accettazione | Il\ campione\ generato\ è\ \mathbf{s}) \\ &= p_2(\mathbf{s})\, \frac{p_1(\mathbf{s})}{B\, p_2(\mathbf{s})} \\ &= \frac{p_1(\mathbf{s})}{B} \end{aligned} \tag{12.62}$$

mentre la probabilità di generare un campione qualsiasi e di accettarlo in un ciclo è pari a

$$\begin{aligned} &Pr(Generare\ un\ campione\ immediatamente\ accettato) \\ &= \sum_{\mathbf{s}\in\mathcal{S}} Pr(Generare\ il\ campione\ \mathbf{s}\ immediatamente\ accettato) \\ &= \sum_{\mathbf{s}\in\mathcal{S}} \frac{p_1(\mathbf{s})}{B} \\ &= \frac{1}{B}. \end{aligned} \tag{12.63}$$

Come immediata estensione, è facile vedere che la probabilità che siano necessari n cicli per generare e accettare un dato campione $\mathbf{s}$ è

$$\begin{aligned} &Pr(Necessari\ n\ cicli\ per\ generare\ e\ accettare\ il\ campione\ \mathbf{s}) \\ &= Pr(Rifiuti\ nei\ primi\ n-1\ cicli) \\ &\times\ Pr(Al\ ciclo\ n-mo\ si\ genera\ il\ campione\ \mathbf{s}\ immediatamente\ accettato) \\ &= \left(1 - \frac{1}{B}\right)^{n-1} \frac{p_1(\mathbf{s})}{B}. \end{aligned} \tag{12.64}$$

Come conseguenza di quanto scritto fino ad ora, la probabilità di generare con lo schema di rigetto è data da

$$\begin{aligned} & Pr(Si\ genera\ il\ campione\ \mathbf{s}) \\ &= \sum_{n=1}^{\infty} Pr(Necessari\ n\ cicli\ per\ generare\ e\ accettare\ il\ campione\ \mathbf{s}) \\ &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{B}\right)^{n-1} \frac{p_1(\mathbf{s})}{B} \\ &= p_1(\mathbf{s}) \end{aligned} \tag{12.65}$$

ossia il campione $\mathbf{s}$ può effettivamente essere considerato come selezionato tramite il disegno $(\mathcal{S}, p_1(\cdot))$.

La probabilità di generare e accettare un campione in un ciclo è una misura molto importante dell'efficienza numerica del metodo di rigetto, poiché quanto più alta è tale probabilità, tanto più rapido è l'algoritmo a generare un campione con le caratteristiche desiderate. In genere il termine $p_1(\mathbf{s})/(B\,p_2(\mathbf{s}))$ è detto *tasso di accettazione condizionato* (*conditional acceptance rate*) ed è indicato con il simbolo $CAR(\mathbf{s})$, mentre $1/B$ è il *tasso di accettazione non condizionato* (*conditional acceptance rate*), ed è indicato con il simbolo AR.

Una caratteristica assai importante dello schema di rigetto, e che sarà utilmente sfruttata nel Capitolo 15 per selezionare campioni mediante un disegno di Poisson condizionato, è che per essere messo in pratica non è necessario conoscere esattamente le probabilità $p_1(\mathbf{s})$. Ad esempio, potrebbe essere $p_1(\mathbf{s}) = c\,q(\mathbf{s})$, con c costante non necessariamente nota esplicitamente, e con le $q(\mathbf{s})$ positive ma non necessariamente aventi somma 1. In questo caso (Esercizio 12.13) l'algoritmo di rigetto si applica esattamente come nei Passi 1-3, purché sia nota una costante B tale che $q(\mathbf{s}) \leqslant B\,p_2(\mathbf{s})$, e purché al Passo 2 l'accettazione di $\mathbf{s}$ avvenga se $U \leqslant q(\mathbf{s})/(B\,p_2(\mathbf{s}))$.

Esercizi

12.1. Consideriamo una popolazione finita di ampiezza $N = 4$, $I_4 = \{1, 2, 3, 4\}$. Supponiamo poi che lo spazio dei campioni sia formato dai seguenti quattro campioni:

$$\mathbf{s}_1 = (1,\, 2),\ \mathbf{s}_2 = (1,\, 3),\ \mathbf{s}_3 = (2,\, 3),\ \mathbf{s}_4 = (1,\, 2,\, 3); \mathbf{s}_5 = (1,\, 3,\, 4)$$

con le seguenti probabilità

$$p(\mathbf{s}_1) = 0.25,\ p(\mathbf{s}_2) = 0.3,\ p(\mathbf{s}_3) = 0.2,\ p(\mathbf{s}_4) = 0.1,\ p(\mathbf{s}_5) = 0.15.$$

a. Calcolare la probabilità di inclusione di primo e secondo ordine per tutte le unità della popolazione.

b. Calcolare $E(n_{\mathbf{s}})$ nei seguenti modi:
   – utilizzando la definizione diretta;
   – utilizzando la formula che esprime $E(n_{\mathbf{s}})$ in funzione delle probabilità di inclusione di primo ordine.

12.2. Consideriamo una popolazione finita U di dimensione N e sia $p(\mathbf{s})$ un disegno il cui spazio campionario costituito da $N + 1$ campioni risulta così cosituito:

– N campioni di dimensione pari a 1, ogni campione contiene una unità della popolazione;
– 1 campione di dimensione N contenente tutte le unità della popolazione.

Supponiamo inoltre che il disegno campionario sia equiprobabile.

a. Qual è la probabilità di inclusione della generica unità i?
b. Qual è la dimensione campionaria attesa?

12.3. Data una popolazione finita di $N = 6$ unità, $I_6 = \{1, 2, 3, 4, 5, 6\}$, si consideri il disegno campionario determinato dallo schema seguente:

– si seleziona una delle tre unità 1, 2, 3, rispettivamente con probabilità 0.5, 0.25, 0.25;
– se si seleziona l'unità 1, il campione è $\mathbf{s}_1 = \{1, 6\}$; se si seleziona l'unità 2, il campione è $\mathbf{s}_2 = \{2, 5\}$; se si seleziona l'unità 3, il campione è $\mathbf{s}_2 = \{3, 4\}$.

Descrivere lo spazio dei campioni $\mathcal{S}$ e le probabilità dei campioni $\mathbf{s} \in \mathcal{S}$. Calcolare inoltre le probabilità di inclusione del primo e del secondo ordine delle unità della popolazione.

12.4. Data una popolazione finita di $N = 5$ unità, $I_5 = \{1, 2, 3, 4, 5\}$, si considerino i 5 numeri

$$p_1 = 0.1, p_2 = 0.2, p_3 = 0.3, p_4 = 0.2, p_5 = 0.2.$$

Si consideri poi il disegno campionario definito dal seguente schema:

– si seleziona una delle cinque unità 1, 2, 3, 4, 5 rispettivamente con probabilità p_1, p_2, p_3, p_4, p_5;
– se al passo 1 si è selezionata l'unità i, si seleziona l'unità $j \neq i$ con probabilità $p_j/(1 - p_i)$.

Descrivere lo spazio dei campioni $\mathcal{S}$ e le probabilità dei campioni $\mathbf{s} \in \mathcal{S}$. Calcolare inoltre le probabilità di inclusione del primo e del secondo ordine delle unità della popolazione.

12.5. Si consideri ancora la popolazione finita di $N = 5$ unità dell'Esercizio 12.4, e i 5 numeri

$$p_1 = 0.1, p_2 = 0.2, p_3 = 0.3, p_4 = 0.2, p_5 = 0.2.$$

Si consideri poi il disegno campionario definito dal seguente schema:

- si seleziona una delle cinque unità 1, 2, 3, 4, 5 rispettivamente con probabilità p_1, p_2, p_3, p_4, p_5;
- se al passo 1 si è selezionata l'unità i, dalla popolazione restante $I_5 \setminus \{i\}$ si seleziona un campione semplice senza ripetizione di due unità distinte.

Le unità selezionate sono quindi tre, tutte distinte. Descrivere lo spazio dei campioni $\mathcal{S}$ e le probabilità dei campioni $\mathbf{s} \in \mathcal{S}$. Calcolare inoltre le probabilità di inclusione del primo e del secondo ordine delle unità della popolazione.

12.6. Data una popolazione finita di $N = 7$ unità, $I_7 = \{1, 2, 3, 4, 5, 6, 7\}$, si considerino i 7 numeri

$$\pi_1^0 = 0.3,\ \pi_2^0 = 0.2,\ \pi_3^0 = 0.5,\ \pi_4^0 = 0.2,\ \pi_5^0 = 0.4, \pi_6^0 = 0.3,\ \pi_7^0 = 0.1.$$

Si consideri poi il disegno campionario definito dal seguente schema:

- si genera un numero aleatorio U con distribuzione uniforme in $[0, 1]$;
- se $0 \leqslant U \leqslant \pi_1^0$ si seleziona l'unità 1; se $\pi_1^0 < U \leqslant \pi_1^0 + \pi_2^0$ si seleziona l'unità 2; se $\pi_1^0 + \pi_2^0 < U \leqslant \pi_1^0 + \pi_2^0 + \pi_3^0$ si seleziona l'unità 3;
- se $1 < U + 1 \leqslant 1 + \pi_4^0$ si seleziona l'unità 4; se $1 + \pi_4^0 < U + 1 \leqslant 1 + \pi_4^0 + \pi_5^0$ si seleziona l'unità 5; se $1 + \pi_4^0 + \pi_5^0 < U + 1 \leqslant 1 + \pi_4^0 + \pi_5^0 + \pi_6^0$ si seleziona l'unità 6; se $1 + \pi_4^0 + \pi_5^0 + \pi_6^0 < U + 1 \leqslant 1 + \pi_4^0 + \pi_5^0 + \pi_6^0 + \pi_7^0$ si seleziona l'unità 7.

Ciascun campione è evidentemente composto da $n = 2$ unità. Calcolare le probabilità di inclusione del primo e del secondo ordine.

12.7. Dato un disegno *ppswor* di numerosità $n = 2$, provare che:

$$\pi_i = p_i \left\{ 1 + \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{p_j}{1 - p_j} \right\} \quad i = 1, \ldots, N;$$

$$\pi_{ij} = p_i p_j \frac{2 - p_i - p_j}{(1 - p_i)(1 - p_j)} \quad i \neq j = 1, \ldots, N.$$

12.8. Provare che nel disegno di Midzuno-Lahiri (riduzione dello schema della Sezione 12.4.3) lo spazio dei campioni è $\mathcal{S} = \mathcal{C}_{N,n}$, e ciascun campione ha probabilità

$$p(\mathbf{s}) = \frac{1}{\binom{N-1}{n-1}} \sum_{i \in \mathbf{s}} p_i, \quad \mathbf{s} \in \mathcal{C}_{N,n}.$$

Suggerimento. La combinazione senza ripetizione $\mathbf{s}$ è ottenuta estraendo nella prima prova una qualsiasi unità $i \in \mathbf{s}$, e nelle altre $n - 1$ prove le restanti $n - 1$ unità di $\mathbf{s}$. Quindi: $p(\mathbf{s}) = \sum_{i \in \mathbf{s}} p_i \frac{1}{\binom{N-1}{n-1}}$.

12.9. Provare la disuguaglianza (12.44).

12.10. Dato un campione $\mathbf{s}$ formato con l'algoritmo sequenziale, provare la relazione (12.57).

12.11. Verificare che se (1) $\mathcal{S}_1$ è l'insieme delle disposizioni con ripetizione di n unità della popolazione e ogni disposizione $\mathbf{s} \in \mathcal{S}_1$ ha probabilità $1/\{N\,(N-1)\cdots(N-n+1)\}$, e (2) $(\mathcal{S}_2,\, p_2(\cdot))$ è il disegno semplice con ripetizione, vale la relazione (12.58).
Suggerimento. Si ha $p_2(\mathbf{s}) = 1/N^n$, e $\sum_{\mathbf{s}\in\mathcal{S}_1} p_2(\mathbf{s}) = N\,(N-1)\cdots(N-n+1)/N^n$.

12.12. Verificare che se $(\mathcal{S}_1,\, p_1(\cdot))$ è il disegno *ppswr* e $(\mathcal{S}_2,\, p_2(\cdot))$ è il disegno *ppswor*, vale la relazione (12.58).

12.13. Si supponga $p_1(\mathbf{s}) = c\,q(\mathbf{s})$, con c costante non necessariamente nota esplicitamente, e con le $q(\mathbf{s})$ positive ma non necessariamente aventi somma 1. Si assuma anche che vi sia nota una costante B tale che $q(\mathbf{s}) \leqslant B\,p_2(\mathbf{s})$. Provare che se al Passo 2 dell'algoritmo di rigetto l'accettazione di $\mathbf{s}$ avviene se $U \leqslant q(\mathbf{s})/(B\,p_2(\mathbf{s}))$, l'algoritmo stesso genera campioni da $(\mathcal{S},\, p_1(\cdots))$, con un tasso di accettazione non condizionato $AR = 1/(B\,c)$.

12.14. Sia X una variabile aleatoria, che assume i k valori $x_1,\, x_2,\, \ldots,\, x_k$ rispettivamente con probabilità $p_1,\, p_2,\, \ldots,\, p_k$, e sia $H = -\sum p_j\, \log p_j$ la sua entropia. Provare che H è massima se $p_1 = p_2 = \cdots = p_k = 1/k$.

Suggerimento. Usare la tecnica dei moltiplicatori di Lagrange, con il vincolo $\sum_j p_j = 1$.

12.15. Sia $(\mathcal{S},\; p(\cdot))$ un disegno campionario, e sia $(\mathcal{S}^*,\; p^*(\cdot))$ la sua riduzione. Provare che $(\mathcal{S},\; p(\cdot))$ ha entropia più grande di quella di $(\mathcal{S}^*,\; p^*(\cdot))$.

12.16. Con riferimento alle (12.53), provare le seguenti affermazioni.

a. Al più uno dei numeri c_i può essere uguale a $C/2 + \sqrt{C^2/4 - \pi_i(1-\pi_i)}$.

Suggerimento. Se $c_i = C/2 + \sqrt{C^2/4 - \pi_i(1-\pi_i)}$ e $c_j = C/2 + \sqrt{C^2/4 - \pi_j(1-\pi_j)}$, allora $c_i > C/2$, $c_j > C/2$, da cui l'assurda conclusione $c_i + c_j > C$.

b. Se $d \geqslant 3/4$, non può essere $c_i = C/2 + \sqrt{C^2/4 - \pi_i(1-\pi_i)}$ neanche per un solo indice i.

Suggerimento. Dalla disuguaglianza tra media geometrica e media aritmetica si ha, per $j \neq i$, $\sqrt{C^2/4 - \pi_j(1-\pi_j)} \leqslant C/2 - \pi_j(1-\pi_j)/C$, da cui $c_j \geqslant \pi_j(1-\pi_j)/C$. Sommando rispetto a $j \neq i$ si ha quindi $C - c_i \geqslant (d - 1/4)/c$, da cui, se $d \geqslant 3/4$, $C - c_i \geqslant 1/(2\,C)$. D'altra parte, dalla (12.52) si ha $C - c_i = \pi_i(1-\pi_i)/c_i \leqslant 1/(4\,c_i)$, e quindi $c_i \leqslant C/2$, che contraddice $c_i > C/2$.

13

Principî di base dell'inferenza statistica basata sul disegno campionario*

13.1 La funzione di verosimiglianza

Nella teoria dell'inferenza statistica un ruolo di primo piano è svolto dalla funzione di verosimiglianza. È quindi di un certo interesse studiare se e in che misura il relativo quadro concettuale possa essere adattato al campionamento da popolazioni finite. Prima di iniziare la vera e propria trattazione, è da sottolineare che l'approccio fino ad ora seguito è *basato sul disegno*, nel senso che: (*i*) le modalità y_i non sono generate da alcun modello; (*ii*) l'unica fonte di aleatorietà, di incertezza, è quella dovuta alla selezione del campione di unità, la quale è regolata da un disegno (probabilistico) di campionamento.

Il quadro di riferimento è quello del Capitolo 2. Indichiamo con $\boldsymbol{Y}_N$ il parametro della popolazione (vettore delle modalità etichettate delle unità), e con Ω_N l'insieme dei possibili "valori" di $\boldsymbol{Y}_N$. Al solito, con $(\mathcal{S},\ p(\cdot))$ denotiamo il disegno di campionamento. Con il simbolo $\boldsymbol{y}(\mathbf{s})$ indicheremo invece il campione di modalità etichettate, ossia l'insieme delle coppie $(i,\ y_i)$, per tutte le unità i del campione $\mathbf{s}$. In simboli:

$$\boldsymbol{y}(\mathbf{s}) = \{(i,\ y_i);\ i \in \mathbf{s}\}.$$

Detta poi $r(\mathbf{s})$ la riduzione di $\mathbf{s}$ (insieme delle unità distinte di $\mathbf{s}$, ognuna delle quali compare una sola volta), si indicherà con

$$\boldsymbol{y}(r(\mathbf{s})) = \{(i,\ y_i);\ i \in r(\mathbf{s})\}$$

l'insieme dei dati campionari ridotti. Intuitivamente, $\boldsymbol{y}(r(\mathbf{s}))$ è ottenuto da $\boldsymbol{y}(\mathbf{s})$ togliendo le "cose inutili", ossia le ripetizioni di unità (e delle corrispondenti modalità etichettate) e l'ordine con cui queste sono osservate.

Sia ora $\boldsymbol{Y}'_N = (y'_1 \ldots y'_N)^T$ una qualunque "punto" di Ω_N, ossia uno dei "possibili valori" del parametro della popolazione. Diremo che $\boldsymbol{Y}'_N$ è *compatibile* con i dati campionari $\boldsymbol{y}(\mathbf{s})$ se:

$$y_i = y'_i \ \ \textit{per ogni}\ i \in \mathbf{s} \tag{13.1}$$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_13,

ovvero se per tutte le unità del campione la modalità effettivamente osservata coincide con quella corrispondente del parametro della popolazione.

Diremo invece che $\boldsymbol{Y}'_N$ è *compatibile* con i dati campionari ridotti $\boldsymbol{y}(r(\mathbf{s}))$ se:

$$y_i = y'_i \ \textit{per ogni } i \in r(\mathbf{s}) \tag{13.2}$$

ovvero se per tutte le unità distinte del campione la modalità effettivamente osservata coincide con quella corrispondente del parametro della popolazione.

È evidente che la compatibilità/incompatibilità di $\boldsymbol{Y}'_N$ con $\boldsymbol{y}(\mathbf{s})$ non dipende né dalle ripetizioni (se un'unità compare più volte, ha sempre la stessa modalità), né dall'ordine. Pertanto, $\boldsymbol{Y}'_N$ *è compatibile con* $\boldsymbol{y}(\mathbf{s})$ *se e solo se è compatibile con* $\boldsymbol{y}(r(\mathbf{s}))$.

Esempio 13.1. Si consideri una popolazione di $N = 4$ unità, $I_4 = \{1, 2, 3, 4\}$, da cui si seleziona un campione mediante un disegno $(\mathcal{S}, p(\cdot))$ definito da

$$\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4\}$$
$$\mathbf{s}_1 = (1, 2, 1),\ \mathbf{s}_2 = (2, 1),\ \mathbf{s}_3 = (3, 4),\ \mathbf{s}_4 = (2, 4)$$
$$p(\mathbf{s}_1) = 0.1,\ p(\mathbf{s}_2) = 0.3,\ p(\mathbf{s}_3) = 0.2,\ p(\mathbf{s}_4) = 0.4.$$

Si tratta evidentemente di un disegno con ripetizioni. La sua riduzione è il disegno $(\mathcal{S}^*, p^*(\cdot))$ definito da

$$\mathcal{S}^* = \{\mathbf{s}_1^*, \mathbf{s}_2^*, \mathbf{s}_3^*\}$$
$$\mathbf{s}_1^* = \{1, 2\},\ \mathbf{s}_2^* = \{3, 4\},\ \mathbf{s}_3 = \{2, 4\}$$
$$p^*(\mathbf{s}_1^*) = 0.4,\ p^*(\mathbf{s}_2^*) = 0.2,\ p^*(\mathbf{s}_3^*) = 0.4.$$

Si noti che $r(\mathbf{s}_1) = r(\mathbf{s}_2) = \mathbf{s}_1^*$.

Chiaramente, $\boldsymbol{y}(\mathbf{s}_1)$ è compatibile con $\boldsymbol{Y}'_N$ se e solo se $y'_1 = y_1$, e $y'_2 = y_2$, indipendentemente da quante volte le unità 1, 2 compaiano nel campione. In altre parole, $\boldsymbol{y}(\mathbf{s}_1)$ è compatibile con $\boldsymbol{Y}'_N$ se e solo se $\boldsymbol{y}(\mathbf{s}_1^*)$ è compatibile con $\boldsymbol{Y}'_N$. □

Se si indica con $c(\boldsymbol{y}(\mathbf{s}), \boldsymbol{Y}'_N)$ l'indicatore di compatibilità/incompatibilità di $\boldsymbol{Y}'_N$ con i dati campionari $\boldsymbol{y}(\mathbf{s})$:

$$c(\boldsymbol{y}(\mathbf{s}), \boldsymbol{Y}'_N) = \begin{cases} 1 \textit{ se } \boldsymbol{Y}'_N \textit{ è compatibile con } \boldsymbol{y}(\mathbf{s}) \\ 0 \textit{ altrimenti} \end{cases} \tag{13.3}$$

e con $c(\boldsymbol{y}(r(\mathbf{s})), \boldsymbol{Y}'_N)$ l'indicatore di compatibiltà/incompatibiltà di $\boldsymbol{Y}'_N$ con i dati campionari ridotti $\boldsymbol{y}(r(\mathbf{s}))$:

$$c(\boldsymbol{y}(r(\mathbf{s})), \boldsymbol{Y}'_N) = \begin{cases} 1 \textit{ se } \boldsymbol{Y}'_N \textit{ è compatibile con } \boldsymbol{y}(r(\mathbf{s})) \\ 0 \textit{ altrimenti} \end{cases} \tag{13.4}$$

è immediato verificare, come conseguenza di quanto sopra detto, che i due indicatori $c(\boldsymbol{y}(\mathbf{s}), \boldsymbol{Y}'_N)$ e $c(\boldsymbol{y}(r(\mathbf{s})), \boldsymbol{Y}'_N)$ *coincidono*:

$$c(\boldsymbol{y}(\mathbf{s}), \boldsymbol{Y}'_N) = c(\boldsymbol{y}(r(\mathbf{s})), \boldsymbol{Y}'_N) \ \textit{qualunque sia } \boldsymbol{Y}'_N \in \Omega_N. \tag{13.5}$$

La *funzione di verosimiglianza*, che ha come argomento il parametro $\boldsymbol{Y}'_N$, è definita come la probabilità di osservare i dati campionari $\boldsymbol{y}(\mathbf{s})$ quando il parametro della popolazione è $\boldsymbol{Y}'_N$. In simboli:

$$L(\boldsymbol{Y}'_N) = Pr(\boldsymbol{y}(\mathbf{s});\, \boldsymbol{Y}'_N);\;\; \boldsymbol{Y}'_N \in \Omega_N. \tag{13.6}$$

Ora, se $\boldsymbol{Y}'_N$ è compatibile con i dati $\boldsymbol{y}(\mathbf{s})$, questi sono osservati se e solo se dalla popolazione è selezionato il campione $\mathbf{s}$. Se $\boldsymbol{Y}'_N$ non è compatibile con i dati $\boldsymbol{y}(\mathbf{s})$, questi non possono in nessun caso essere osservati. Si ha quindi

$$\begin{aligned} Pr(\boldsymbol{y}(\mathbf{s});\, \boldsymbol{Y}'_N) &= \begin{cases} p(\mathbf{s}) \textit{ se } \boldsymbol{Y}'_N \textit{ è compatibile con } \boldsymbol{y}(\mathbf{s}) \\ 0 \textit{ altrimenti} \end{cases} \\ &= p(\mathbf{s})\, c(\boldsymbol{y}(\mathbf{s}),\, \boldsymbol{Y}'_N). \end{aligned} \tag{13.7}$$

Usando (13.6) e (13.7), la funzione di verosimiglianza si può quindi scrivere come:

$$L(\boldsymbol{Y}'_N) = p(\mathbf{s})\, c(\boldsymbol{y}(\mathbf{s}),\, \boldsymbol{Y}'_N);\;\; \boldsymbol{Y}'_N \in \Omega_N. \tag{13.8}$$

La (13.8) mette in evidenza un fatto molto importante: la funzione di verosimiglianza assume solo due valori: uno ($p(\mathbf{s})$) per tutti i possibili $\boldsymbol{Y}'_N$ compatibili con i dati $\boldsymbol{y}(\mathbf{s})$, e l'altro (0) per tutti i possibili $\boldsymbol{Y}'_N$ non compatibili con i dati $\boldsymbol{y}(\mathbf{s})$. Quindi, la funzione di verosimiglianza discrimina i parametri $\boldsymbol{Y}'_N$ compatibili con i dati campionari da quelli non compatibili, ma non discrimina tra i diversi parametri $\boldsymbol{Y}'_N$ compatibili con i dati $\boldsymbol{y}(\mathbf{s})$ campionari. La verosimiglianza $L(\boldsymbol{Y}'_N)$ ha quindi una forma *piatta*, in quanto rappresenta come ugualmente verosimili tutti i possibili parametri $\boldsymbol{Y}'_N$ della popolazione compatibili con i dati $\boldsymbol{y}(\mathbf{s})$ osservati a livello campionario. Questa forma rende praticamente inutile $L(\boldsymbol{Y}'_N)$ per fini di inferenza statistica. In termini un pò diversi, ma equivalenti, la forma piatta della funzione di verosimiglianza mostra un fatto pressoché scontato: in assenza di ipotesi aggiuntive, l'osservare le modalità (etichettate) delle unità campionarie non dice nulla sulle unità che non fanno parte del campione. La forma piatta della funzione di verosimiglianza è conseguenza di due fattori:

1. il disegno campionario è *non informativo*, nel senso che le probabilità $p(\mathbf{s})$ dei campioni dipendono dalle unità che li compongono, ma non dalle corrispondenti modalità y_i;
2. i dati campionari sono raccolti nella forma di modalità etichettate, nel senso che viene conservata l'informazione relativa alle unità a cui si riferiscono le modalità osservate in $\boldsymbol{y}(\mathbf{s})$.

Se venisse meno anche uno solo degli elementi 13.1, 13.1, verrebbe anche meno la (13.8). Per ulteriori approfondimenti sulla funzione di verosimiglianza nel campionamento da popolazioni finite si rinvia al volume Cassel *e altri* (1977).

13.2 Sufficienza e minimalità

13.2.1 Statistiche sufficienti

Come già visto nel Capitolo 2, una statistica campionaria (statistica, per brevità) $T = t(\boldsymbol{y}(\mathbf{s}))$ è una qualunque funzione dei dati campionari. Se si "dimenticano" i dati campionari $\boldsymbol{y}(\mathbf{s})$, e si "ricorda" solo il valore $t(\boldsymbol{y}(\mathbf{s}))$ della statistica T, si effettua ovviamente un "riassunto" dei dati stessi.

Intuitivamente, ogni procedura di inferenza statistica dovrebbe basarsi su (almeno) due principi di base.

1. Ci si dovrebbe basare, per ragioni di economicità e sinteticità, su un *riassunto* dei dati campionari, ossia su una opportuna statistica. Quanto più sintetico è il riassunto, tanto meglio è soddisfatta questa esigenza.
2. Il riassunto dei dati campionari $\boldsymbol{y}(\mathbf{s})$ (ossia la statistica su di essi calcolata) dovrebbe conservare tutta l'informazione che essi forniscono sul parametro della popolazione.

Le statistiche che soddisfano i requisiti 13.2.1, 13.2.1 sono dette *statistiche sufficienti*. In via intuitiva, una statistica $T = t(\boldsymbol{y}(\mathbf{s}))$ è *sufficiente* se riassume tutta l'informazione che i dati campionari sono in grado di fornire sul parametro della popolazione. In altre parole, T è sufficiente se una volta noto il valore $t(\boldsymbol{y}(\mathbf{s}))$, la conoscenza dei dati campionari $\boldsymbol{y}(\mathbf{s})$ non fornisce nessuna informazione aggiuntiva sul parametro della popolazione. Ora, i dati campionari $\boldsymbol{y}(\mathbf{s})$ forniscono informazioni sul parametro della popolazione solo perché la loro distribuzione di probabilità dipende dal parametro stesso, come si vede dalla (13.7). Quindi, in termini formali, una statistica T è sufficiente se la distribuzione di probabilità dei dati campionari $\boldsymbol{y}(\mathbf{s})$, condizionata al valore $t(\boldsymbol{y}(\mathbf{s}))$ assunto da T, non dipende dal parametro della popolazione. In simboli:

$$\begin{aligned} &T \textit{ sufficiente significa che } Pr(\boldsymbol{y}(\mathbf{s}) \,|T = t(\boldsymbol{y}(\mathbf{s}));\\ &\qquad\qquad \boldsymbol{Y}'_N) \textit{ non dipende da } \boldsymbol{Y}'_N. \end{aligned} \tag{13.9}$$

Un risultato di base per riconoscere statistiche sufficienti è il *teorema di fattorizzazione di Fisher-Neyman* (cfr. Cox e Hinkley (1974)). *Una statistica $T = t(\boldsymbol{y}(\mathbf{s}))$ è sufficiente (per il parametro della popolazione) se e solo se la probabilità $Pr(\boldsymbol{y}(\mathbf{s});\ \boldsymbol{Y}'_N)$ dei dati campionari può essere fattorizzata nel prodotto (i) di una funzione che dipende sono dal valore $t(\boldsymbol{y}(\mathbf{s}))$ di T e dal parametro $\boldsymbol{Y}'_N$ della popolazione per (ii) una funzione che non dipende da $\boldsymbol{Y}'_N$. In simboli:*

$$Pr(\boldsymbol{y}(\mathbf{s});\ \boldsymbol{Y}'_N) = g(t(\boldsymbol{y}(\mathbf{s}));\ \boldsymbol{Y}'_N)\, h(\boldsymbol{y}(\mathbf{s})). \tag{13.10}$$

Proposizione 13.1. *La riduzione $\boldsymbol{y}(r(\mathbf{s}))$ dei dati campionari è una statistica sufficiente per il parametro $\boldsymbol{Y}'_N$ della popolazione.*

Dimostrazione. In primo luogo, dalla (13.7) e (13.5) si ha

$$\begin{aligned} Pr(\boldsymbol{y}(\mathbf{s});\ \boldsymbol{Y}'_N) &= p(\mathbf{s})\, c(\boldsymbol{y}(\mathbf{s}),\ \boldsymbol{Y}'_N) \\ &= p(\mathbf{s})\, c(\boldsymbol{y}(r(\mathbf{s})),\ \boldsymbol{Y}'_N). \end{aligned} \tag{13.11}$$

Basta a questo punto usare il teorema di fattorizzazione di Fisher-Neyman ponendo nella (13.10) $g(t(\boldsymbol{y}(\mathbf{s}));\ \boldsymbol{Y}'_N) = c(\boldsymbol{y}(r(\mathbf{s})),\ \boldsymbol{Y}'_N)$ e $h(\boldsymbol{y}(\mathbf{s})) = p(\mathbf{s})$. □

13.2.2 In che misura una statistica riassume i dati campionari? Partizioni indotte da statistiche

Data una statistica $T = t(\boldsymbol{y}(\mathbf{s}))$, sia $\mathcal{T} = \{t(\boldsymbol{y}(\mathbf{s}));\ \mathbf{s} \in \mathcal{S}\}$ l'insieme dei valori che può assumere al variare del campione $\mathbf{s}$ in $\mathcal{S}$ (e fissati $y_1, \ldots, y_N$). Per ciascuno "valore possibile" $t \in \mathcal{T}$ indichiamo con $T^{-1}(t)$ l'insieme dei dati campionari $\boldsymbol{y}(\mathbf{s})$ che "producono" il valore t. In simboli:

$$T^{-1}(t) = \{\boldsymbol{y}(\mathbf{s}) :\ t(\boldsymbol{y}(\mathbf{s})) = t\}.$$

Ovviamente, ciascun $T^{-1}(t)$ è un sottoinsieme dello *spazio dei campioni di modalità etichettate* $\boldsymbol{y}(\mathcal{S}) = \{\boldsymbol{y}(\mathbf{s});\ \mathbf{s} \in \mathcal{S}\}$. La famiglia di insiemi (sottoinsiemi di $\boldsymbol{y}(\mathcal{S})$)

$$\begin{aligned} \mathcal{P}_T &= \{T^{-1}(t);\ t \in \mathcal{T}\} \\ &= \{\{\boldsymbol{y}(\mathbf{s}) :\ t(\boldsymbol{y}(\mathbf{s})) = t\};\ t \in \mathcal{T}\} \end{aligned}$$

è la *partizione di* $\boldsymbol{y}(\mathcal{S})$ *indotta da* T. In Fig. 13.1 è rappresentata graficamente una partizione indotta.

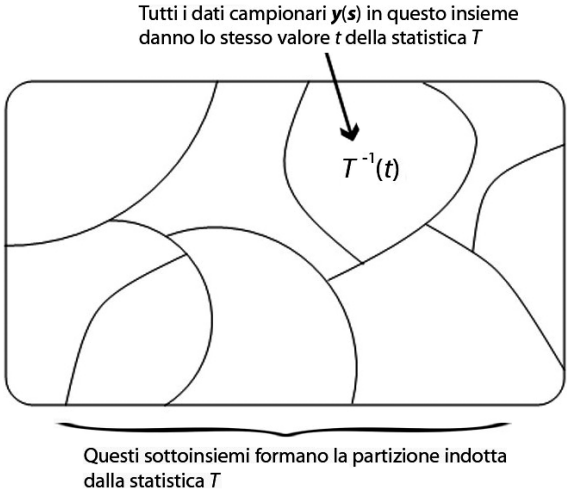


Fig. 13.1 Partizione indotta dalla statistica T

Esempio 13.2. Si consideri una popolazione finita $I_3 = \{1, 2, 3\}$ composta da $N = 3$ unità, e sia $(\mathcal{S},\ p(\cdot))$, un disegno campionario (ordinato e con ripetizioni) definito come:

$$\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6\}$$
$$\mathbf{s}_1 = (1,1,2),\ \mathbf{s}_2 = (1,2,1),\ \mathbf{s}_3 = (1,3),\ \mathbf{s}_4 = (2,1,2),\ \mathbf{s}_5 = (3,2),\ \mathbf{s}_6 = (3,1)$$
$$p(\mathbf{s}_1) = 0.2,\ p(\mathbf{s}_2) = 0.1,\ p(\mathbf{s}_3) = 0.1;\ p(\mathbf{s}_4) = 0.1,\ p(\mathbf{s}_5) = 0.3,\ p(\mathbf{s}_6) = 0.2.$$

I campioni di modalità etichettate (dati campionari) sono elencati di seguito:

$$\boldsymbol{y}(\mathbf{s}_1) = ((1, y_1), (1, y_1), (2, y_2)),\ \boldsymbol{y}(\mathbf{s}_2) = ((1, y_1), (2, y_2), (1, y_1)),$$
$$\boldsymbol{y}(\mathbf{s}_3) = ((1, y_1), (3, y_3)),\ \boldsymbol{y}(\mathbf{s}_4) = ((2, y_2), (1, y_1), (2, y_2)),$$
$$\boldsymbol{y}(\mathbf{s}_5) = ((3, y_3), (2, y_2)),\ \boldsymbol{y}(\mathbf{s}_6) = ((3, y_3), (1, y_1)).$$

La statistica T definita da:

$$t(\boldsymbol{y}(\mathbf{s}_1)) = t_1 = ((1, y_1), (2, y_2)); \quad t(\boldsymbol{y}(\mathbf{s}_2)) = t_1 = ((1, y_1), (2, y_2));$$
$$t(\boldsymbol{y}(\mathbf{s}_3)) = t_2 = ((1, y_1), (3, y_3)); \quad t(\boldsymbol{y}(\mathbf{s}_4)) = t_1 = ((1, y_1), (2, y_2));$$
$$t(\boldsymbol{y}(\mathbf{s}_5)) = t_3 = ((3, y_3), (2, y_2)); \quad t(\boldsymbol{y}(\mathbf{s}_6)) = t_4 = ((3, y_3), (1, y_1))$$

induce la partizione di $\boldsymbol{y}(\mathcal{S})$, $\mathcal{P}_T = \{T^{-1}(t_1),\ T^{-1}(t_2),\ T^{-1}(t_3),\ T^{-1}(t_4)\}$ definita da

$$T^{-1}(t_1) = \{\boldsymbol{y}(\mathbf{s}_1),\ \boldsymbol{y}(\mathbf{s}_2),\ \boldsymbol{y}(\mathbf{s}_4)\};\ T^{-1}(t_2) = \{\boldsymbol{y}(\mathbf{s}_3)\};$$
$$T^{-1}(t_3) = \{\boldsymbol{y}(\mathbf{s}_5)\};\ T^{-1}(t_4) = \{\boldsymbol{y}(\mathbf{s}_6)\}.$$

Posto infine, per ciascun $t = t_1, t_2, t_3, t_4$,

$$\gamma(T^{-1}(t),\ \boldsymbol{Y}'_N) = \textit{più grande valore di } c(\boldsymbol{y}(\mathbf{s}),\ \boldsymbol{Y}'_N),\ \textit{con } \boldsymbol{y}(\mathbf{s}) \in T^{-1}(t)$$
$$\zeta(t) = \textit{somma delle } p(\mathbf{s}) \textit{ per tutti i campioni } \mathbf{s} \textit{ tali che } T(\boldsymbol{y}(\mathbf{s}))$$
$$= t$$

è facile vedere che vale la relazione

$$Pr(T = t;\ \boldsymbol{Y}'_N) = \gamma(T^{-1}(t),\ \boldsymbol{Y}'_N)\zeta(t)$$

da cui, usando il teorema di fattorizzazione di Fisher-Neyman, segue subito la sufficienza di T. □

La nozione di partizione indotta da una statistica permette di capire con facilità il significato dell'affermazione: "Una statistica è un riassunto dei dati campionari". Riassumere i dati campionari $\boldsymbol{y}(\mathbf{s})$ con il valore $t(\boldsymbol{y}(\mathbf{s})) = t$ assunto dalla statistica T significa che tutti i "punti" (dati campionari) dell'insieme $T^{-1}(t)$ vengono sostituiti dal valore t. In altre parole, tutti i campioni di modalità etichettate $\boldsymbol{y}(\mathbf{s})$ in $T^{-1}(t)$ sono *equivalenti*, in quanto forniscono lo stesso valore t della statistica T.

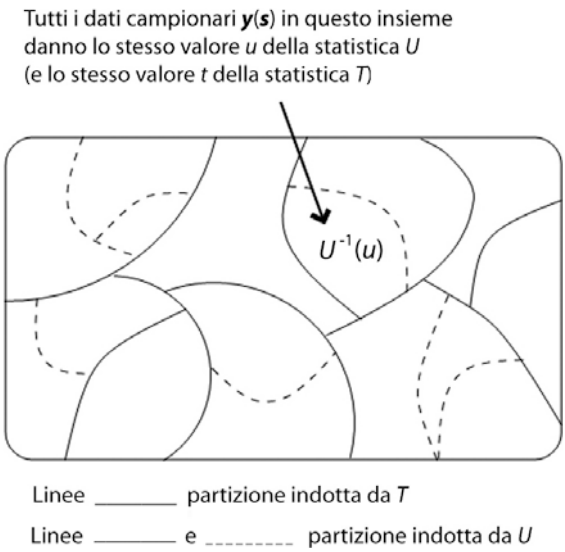


Fig. 13.2 Partizioni indotte dalle statistiche T (meno fine) e U (più fine)

Questa semplice osservazione permette anche di chiarire "in che misura" una statistica riassume i dati campionari. Consideriamo due statistiche T, U, le quali assumono valori rispettivamente in $\mathcal{T}, \mathcal{U}$, e indichiamo con $\mathcal{P}_T$, $\mathcal{P}_U$ le corrispondenti partizioni indotte. Diremo che $\mathcal{P}_U$ è *più fine* di $\mathcal{P}_T$ (o che $\mathcal{P}_T$ è meno fine di $\mathcal{P}_U$) se ogni insieme $U^{-1}(u)$ è contenuto in un insieme $T^{-1}(t)$. In maniera equivalente, si può dire che la partizione indotta da T è meno fine di quella indotta da U se per ogni coppia di campioni (di modalità etichettate) tali che $u(\boldsymbol{y}(\mathbf{s}_1)) = u(\boldsymbol{y}(\mathbf{s}_2))$ si ha anche $t(\boldsymbol{y}(\mathbf{s}_1)) = t(\boldsymbol{y}(\mathbf{s}_2))$. In termini grafici, ciò è illustrato in Fig. 13.2. Ad ogni modo, una statistica riassume tanto più i dati campionari quanto meno fine è la partizione $\boldsymbol{y}(\mathcal{S})$ che induce.

Esempio 13.3. Si consideri ancora l'Esempio 13.2, e si definisca la statistica U come:

$$u(\boldsymbol{y}(\mathbf{s}_1)) = 1; \quad u(\boldsymbol{y}(\mathbf{s}_2)) = 1; \quad u(\boldsymbol{y}(\mathbf{s}_3)) = 2;$$
$$u(\boldsymbol{y}(\mathbf{s}_4)) = 3; \quad u(\boldsymbol{y}(\mathbf{s}_5)) = 4; \quad u(\boldsymbol{y}(\mathbf{s}_6)) = 5.$$

Ragionando come nell'Esempio 13.2, si vede subito che anche U è sufficiente per il parametro della popolazione.

La statistica U induce la partizione di $\boldsymbol{y}(\mathcal{S})$:

$$\mathcal{P}_U = \{U^{-1}(1),\, U^{-1}(2),\, U^{-1}(3),\, U^{-1}(4),\, U^{-1}(5)\}$$

definita da

$$U^{-1}(1) = \{\boldsymbol{y}(\mathbf{s}_1),\, \boldsymbol{y}(\mathbf{s}_2)\};\;\; U^{-1}(2) = \{\boldsymbol{y}(\mathbf{s}_3)\};\;\; U^{-1}(3) = \{\boldsymbol{y}(\mathbf{s}_4)\};$$
$$U^{-1}(4) = \{\boldsymbol{y}(\mathbf{s}_5)\};\;\; U^{-1}(5) = \{\boldsymbol{y}(\mathbf{s}_6)\}.$$

Chiaramente, $\mathcal{P}_U$ è più fine della partizione $\mathcal{P}_T$ dell'Esempio 13.3. Questo ha una conseguenza importante: *la statistica T può essere espressa come funzione di U*. Per rendersene conto basta osservare che:

$$\begin{aligned} T = t_1 \;\; &se \;\; U = 1 \\ T = t_1 \;\; &se \;\; U = 3 \\ T = t_2 \;\; &se \;\; U = 2 \\ T = t_3 \;\; &se \;\; U = 4 \\ T = t_4 \;\; &se \;\; U = 5\,. \end{aligned}$$

Se si definisce quindi la funzione $f(U)$ nel modo seguente:

$$f(u) = \begin{cases} t_1 \; se \; U = 1 \\ t_1 \; se \; U = 3 \\ t_2 \; se \; U = 2 \\ t_3 \; se \; U = 4 \\ t_4 \; se \; U = 5 \end{cases}$$

si vede subito che $T = f(U)$. □

Quanto evidenziato alla fine dell'Es. 13.3 vale del tutto in generale: *se una statistica U induce una partizione più fine di quella indotta dalla statistica T, allora è possibile esprimere T come funzione di U*. Per la dimostrazione di questo semplice fatto si rinvia all'Esercizio 13.1. L'idea della dimostrazione è comunque semplicissima: se U induce una partizione più fine di quella indotta da T, allora T assume lo stesso valore per tutti i campioni (di modalità etichettate) in $U^{-1}(u)$.

La nozione di partizione indotta da una statistica permette anche di chiarire un altro punto importante. Due statistiche U, V, pur essendo differenti in quanto assumono valori diversi, possono indurre la stessa partizione di $\boldsymbol{y}(\mathcal{S})$. In questo caso esse sono *equivalenti*, in quanto riassumono nello stesso modo i dati campionari (vds. Es. 13.4). Ragionando come in precedenza, è chiaro che se $\mathcal{P}_U$ e $\mathcal{P}_V$ coincidono, è possibile sia esprimere U in funzione di V che V in funzione di U. Pertanto, U e V sono in *corrispondenza biunivoca*. In simboli:

$$\mathcal{P}_U = \mathcal{P}_V \;\; \textit{se solo se } U \textit{ e } V \textit{ sono in corrispondenza biunivoca.}$$

Esempio 13.4. Si consideri ancora l'Esempio 13.3, e si definisca la statistica V come:

$$\begin{gathered} v(\boldsymbol{y}(\mathbf{s}_1)) = 100; \quad v(\boldsymbol{y}(\mathbf{s}_2)) = 100; \quad v(\boldsymbol{y}(\mathbf{s}_3)) = 10; \\ v(\boldsymbol{y}(\mathbf{s}_4)) = 5; \quad v(\boldsymbol{y}(\mathbf{s}_5)) = 9; \quad v(\boldsymbol{y}(\mathbf{s}_6)) = 4. \end{gathered}$$

La partizione di $\boldsymbol{y}(\mathcal{S})$ indotta da V,

$$\mathcal{P}_V = \{V^{-1}(4),\; V^{-1}(5),\; U^{-1}(9),\; U^{-1}(10),\; U^{-1}(100)\}$$

è definita da

$$V^{-1}(4) = \{\boldsymbol{y}(\mathbf{s}_6)\};\ \ V^{-1}(5) = \{\boldsymbol{y}(\mathbf{s}_4)\};\ \ V^{-1}(9) = \{\boldsymbol{y}(\mathbf{s}_5)\};$$
$$V^{-1}(10) = \{\boldsymbol{y}(\mathbf{s}_3)\};\ \ V^{-1}(100) = \{\boldsymbol{y}(\mathbf{s}_1),\ \boldsymbol{y}(\mathbf{s}_2)\}.$$

Essa coincide con la partizione indotta da U: $\mathcal{P}_V = \mathcal{P}_U$.

È infine immediato verificare che U e V sono in corrispondenza biunivoca. Basta definire la funzione $f(U)$ come:

$$f(u) = \begin{cases} 100\ se\ u = 1 \\ 10\ se\ u = 2 \\ 5\ se\ u = 3 \\ 9\ se\ u = 4 \\ 4\ se\ u = 5 \end{cases}$$

e verificare che $V = f(U)$, $U = f^{-1}(V)$. □

13.2.3 Statistiche sufficienti minimali

La proprietà di sufficienza può facilmente essere enunciata in termini di partizione indotta. Infatti, la (13.9) equivale a dire che T è sufficiente se e solo se la distribuzione di probabilità dei dati campionari "ristretta" a ciascun insieme $T^{-1}(t)$ non dipende dal parametro $\boldsymbol{Y}'_N$ della popolazione.

La nozione di partizione indotta permette anche di risolvere un importante problema. In generale, vi sono più statistiche sufficienti per il parametro della popolazione (si vedano in proposito gli Esempi 13.2, 13.3). Sorge quindi il problema di *quale* di esse scegliere per riassumere i dati. Coerentemente con quanto detto nelle sezioni precedenti, risulta naturale scegliere la statistica sufficiente che riassume "il più possibile" i dati campionari. In termini un po' più formali, l'idea è quella di scegliere la statistica sufficiente che induce la partizione *meno fine* di $\boldsymbol{y}(\mathcal{S})$. Essa è detta *statistica sufficiente minimale*. Chiaramente, questo equivale a dire che la statistica sufficiente minimale è una statistica sufficiente che può essere espressa come funzione di ogni altra statistica sufficiente. A sua volta, questo equivale a dire che una statistica sufficiente $T = t(\boldsymbol{y}(\mathbf{s}))$ è minimale se per ogni altra statistica sufficiente $U = u(\boldsymbol{y}(\mathbf{s}))$ la relazione $u(\boldsymbol{y}(\mathbf{s}_1)) = u(\boldsymbol{y}(\mathbf{s}_2))$ implica che $t(\boldsymbol{y}(\mathbf{s}_1)) = t(\boldsymbol{y}(\mathbf{s}_2))$.

Proposizione 13.2. *La riduzione $\boldsymbol{y}(r(\mathbf{s}))$ dei dati campionari è una statistica sufficiente minimale per il parametro $\boldsymbol{Y}'_N$ della popolazione.*

Dimostrazione. La sufficienza di $\boldsymbol{y}(r(\mathbf{s}))$ è già stata dimostrata. Per provare la sua minimalità basta mostrare che se $T = t(\boldsymbol{y}(\mathbf{s}))$ è una qualunque altra statistica sufficiente, e se per due campioni $\mathbf{s}_1$, $\mathbf{s}_2$ si ha $t(\boldsymbol{y}(\mathbf{s}_1)) = t(\boldsymbol{y}(\mathbf{s}_2))$, allora è anche $\boldsymbol{y}(r(\mathbf{s}_1)) = \boldsymbol{y}(r(\mathbf{s}_2))$.

Se T è sufficiente per $\boldsymbol{Y}'_N$, si ha anzitutto, dal teorema di fattorizzazione di Fisher-Neyman, che

$$Pr(\boldsymbol{y}(\mathbf{s}_1);\ \boldsymbol{Y}'_N) = g(t(\boldsymbol{y}(\mathbf{s}_1));\ \boldsymbol{Y}'_N)\, h(\boldsymbol{y}(\mathbf{s}_1)) \tag{13.12}$$

e similmente, essendo anche $t(\boldsymbol{y}(\mathbf{s}_1)) = t(\boldsymbol{y}(\mathbf{s}_2))$,

$$\begin{aligned} Pr(\boldsymbol{y}(\mathbf{s}_2);\ \boldsymbol{Y}'_N) &= g(t(\boldsymbol{y}(\mathbf{s}_2));\ \boldsymbol{Y}'_N)\, h(\boldsymbol{y}(\mathbf{s}_2)) \\ &= g(t(\boldsymbol{y}(\mathbf{s}_1));\ \boldsymbol{Y}'_N)\, h(\boldsymbol{y}(\mathbf{s}_2)). \end{aligned} \tag{13.13}$$

Si osservi che $h(\boldsymbol{y}(\mathbf{s}_1)) > 0$, $h(\boldsymbol{y}(\mathbf{s}_2)) > 0$, perché in caso contrario il membro di destra della (13.12) e della (13.13) sarebbe identicamente nullo, per tutti i possibili valori $\boldsymbol{Y}'_N$. Da ciò discende l'uguaglianza

$$\frac{Pr(\boldsymbol{y}(\mathbf{s}_2);\ \boldsymbol{Y}'_N)}{h(\boldsymbol{y}(\mathbf{s}_2))} = \frac{Pr(\boldsymbol{y}(\mathbf{s}_1);\ \boldsymbol{Y}'_N)}{h(\boldsymbol{y}(\mathbf{s}_1))}$$

da cui segue che

$$Pr(\boldsymbol{y}(\mathbf{s}_2);\ \boldsymbol{Y}'_N) = Pr(\boldsymbol{y}(\mathbf{s}_1);\ \boldsymbol{Y}'_N) \frac{h(\boldsymbol{y}(\mathbf{s}_2))}{h(\boldsymbol{y}(\mathbf{s}_1))}$$

e quindi, usando la (13.11),

$$p(\mathbf{s}_1)\, c(\boldsymbol{y}(r(\mathbf{s}_2)),\, \boldsymbol{Y}'_N) = p(\mathbf{s}_2)\, c(\boldsymbol{y}(r(\mathbf{s}_1)),\, \boldsymbol{Y}'_N)\, \frac{h(\boldsymbol{y}(\mathbf{s}_2))}{h(\boldsymbol{y}(\mathbf{s}_1))}. \tag{13.14}$$

La (13.14) mostra che $c(\boldsymbol{y}(r(\mathbf{s}_1)),\, \boldsymbol{Y}'_N) = 1$ ogni volta che $c(\boldsymbol{y}(r(\mathbf{s}_2)),\, \boldsymbol{Y}'_N) = 1$, e $c(\boldsymbol{y}(r(\mathbf{s}_1)),\, \boldsymbol{Y}'_N) = 0$ ogni volta che $c(\boldsymbol{y}(r(\mathbf{s}_2)),\, \boldsymbol{Y}'_N) = 0$. In altre parole, deve essere

$$c(\boldsymbol{y}(r(\mathbf{s}_1)),\, \boldsymbol{Y}'_N) = c(\boldsymbol{y}(r(\mathbf{s}_2)),\, \boldsymbol{Y}'_N) \ \ \textit{qualunque sia}\ \boldsymbol{Y}'_N \in \Omega_N. \tag{13.15}$$

Dalla (13.15) è facile desumere che $\boldsymbol{y}(r(\mathbf{s}_1))$ coincide con $\boldsymbol{y}(r(\mathbf{s}_2))$, e questo completa la dimostrazione. □

Le Proposizioni 13.1, 13.2 mettono in evidenza due fatti importantissimi:

1. *le ripetizioni e l'ordine non danno nessuna informazione aggiuntiva rispetto a quella fornita dalle unità campionarie distinte e dalle corrispondenti modalità etichettate;*
2. *la riduzione dei dati campionari è il massimo riassunto dei dati campionari stessi che non fa perdere informazione.*

L'affermazione 13.2.3, già introdotta in via intuitiva nel Capitolo 2, trova la sua giustificazione formale nella Proposizione 13.1. L'affermazione 13.2.3, invece, è una diretta conseguenza della Proposizione 13.2.

Un'ultima osservazione prima di concludere. La proprietà di sufficienza, ed in particolare quella di sufficienza minimale, di una statistica T *dipendono dalla partizione* $\mathcal{P}_T$ *indotta da* T. In particolare, *ogni statistica* T *che induce la stessa partizione della riduzione dei dati campionari è essa stessa sufficiente minimale* (ed è in corrispondenza biunivoca con $\boldsymbol{y}(r(\mathbf{s}))$).

13.3 Perché bisogna basare l'inferenza su statistiche sufficienti minimali: il teorema di Rao-Blackwell

Il discorso svolto nelle sezioni precedenti mostra che il riassumere i dati campionari mediante la loro riduzione conserva interamente l'informazione fornita dai dati stessi. Risulta quindi naturale basare l'inferenza (in particolare, ma non solo, la stima della media della popolazione) sulla riduzione $\boldsymbol{y}(r(\mathbf{s}))$. A questo punto sorge però una questione. Cosa accade se *non* si riassumono i dati campionari con una statistica sufficiente minimale e ci si basa sui dati originari $\boldsymbol{y}(\mathbf{s})$, eventualmente contenenti ripetizioni e/o ordine? La risposta è fornita dal *teorema di Rao-Blackwell*, di seguito enunciato e provato.

Proposizione 13.3. (Teorema di Rao-Blackwell) *Sia $\widehat{\theta}$ uno stimatore del parametro $\theta = \theta(\boldsymbol{Y}_N)$, e sia T una statistica sufficiente per il parametro della popolazione. Posto:*

$$\widehat{\theta}^* = E[\widehat{\theta}\,|T] \tag{13.16}$$

valgono le tre seguenti relazioni

$$E[\widehat{\theta}^*] = E[\widehat{\theta}]; \tag{13.17}$$

$$V(\widehat{\theta}^*) \leqslant V(\widehat{\theta}); \tag{13.18}$$

$$MSE(\widehat{\theta}^*) \leqslant MSE(\widehat{\theta}). \tag{13.19}$$

In (13.18) *e* (13.19) *il segno $=$ vale se e solo se $\widehat{\theta}^* = \widehat{\theta}$ con probabilità 1. In tutti gli altri casi vale la disuguaglianza stretta $<$.*

Dimostrazione. La (13.17) è una conseguenza immediata di una ben nota proprietà della media condizionata: il valore atteso della media condizionata è uguale al valore atteso non condizionato.

La disuguaglianza (13.18) è una semplice conseguenza della formula di decomposizione della varianza, dalla quale si ha

$$\begin{aligned} V(\widehat{\theta}) &= V(E[\widehat{\theta}\,|T]) + E[V(\widehat{\theta}\,|T)] \\ &= V(\widehat{\theta}^*) + E[V(\widehat{\theta}\,|T)] \end{aligned}$$

e quindi, essendo $V(\widehat{\theta}\,|T) \geqslant 0$, segue la (13.18). In particolare, il segno $=$ vale se e solo se $V(\widehat{\theta}\,|T) = 0$ con probabilità 1, ossia se e solo se $\widehat{\theta} = E[\widehat{\theta}\,|T]$ con probabilità 1.

Infine, la (13.19) segue dalla relazione

$$MSE(\widehat{\theta}) = V(\widehat{\theta}) + (E[\widehat{\theta}] - \theta)^2$$

e dalle (13.17), (13.18). □

Prima di discutere il significato del teorema di Rao-Blackwell, è necessaria qualche osservazione. In primo luogo, la condizione che T sia una statistica sufficiente serve a far sì che il valore atteso condizionato $E[\widehat{\theta}\,|T]$ dipenda *solo* dai dati campionari, e non dall'intero parametro della popolazione. Questo è proprio ciò che accadrebbe se T non fosse sufficiente. In secondo luogo, se la statistica T fosse sufficiente ma non minimale, si potrebbe riapplicare la (13.16), ottenendo un nuovo stimatore migliore di quello di partenza. Quindi, l'unica statistica sufficiente che vale la pena considerare in (13.16) è quella minimale.

Il significato del teorema di Rao-Blackwell è molto semplice. Se uno stimatore $\widehat{\theta}$ non è funzione della statistica sufficiente minimale, si può costruire un suo *miglioramento* calcolando la sua media condizionata rispetto alla statistica sufficiente minimale, come appare in (13.16). L'operazione che porta da $\widehat{\theta}$ a $\widehat{\theta}^*$ è detta *Rao-Blackwellizzazione* di $\widehat{\theta}$. In sostanza, si ha una dicotomia: (*i*) se uno stimatore $\widehat{\theta}$ non è funzione della statistica sufficiente minimale, allora può essere migliorato tramite la (13.16); (*ii*) se lo stimatore $\widehat{\theta}$ è già funzione della statistica sufficiente minimale, allora la (13.16) non porta a nessun miglioramento, in quanto è $\widehat{\theta}^* = \widehat{\theta}$ (l'operazione di Rao-Blackwellizzazione riproduce lo stimatore di partenza). Poiché la statistica sufficiente minimale è null'altro che l'insieme dei dati campionari (etichettati) privati di ripetizione e ordine, la dicotomia si può leggere come segue: (*i*) se uno stimatore $\widehat{\theta}$ dipende dalle ripetizioni e/o dall'ordine con cui le unità sono osservate nel campione, allora può essere migliorato tramite la (13.16); (*ii*) se lo stimatore $\widehat{\theta}$ non dipende né dalle ripetizioni e né dall'ordine, allora è già funzione della statistica sufficiente minimale, e non può essere migliorato tramite la (13.16).

Esempio 13.5. Sia $I_N = \{1, \ldots, N\}$ una popolazione finita di numerosità N, e si consideri un disegno semplice con ripetizione di numerosità n. Lo spazio dei campioni è l'insieme di tutte le n-ple ordinate (disposizioni con ripetizione di classe n) di unità della popolazione: $\mathcal{S} = I_N \times \cdots \times I_N$; ogni campione ha probabilità $1/N^n$. Si noti che questo disegno è *totalmente simmetrico*, nel senso che se si scambiano tra loro (tecnicamente, se si permutano) le etichette delle unità della popolazione, il disegno rimane inalterato. Esattamente come accade per il disegno semplice senza ripetizione, il disegno campionario tratta "alla pari", in modo simmetrico, tutte le unità della popolazione.

Come stimatore della media della popolazione consideriamo la media campionaria

$$\overline{y}_{\mathbf{s}} = \frac{1}{n} \sum_{i \in \mathbf{s}} y_i. \tag{13.20}$$

Lo stimatore $\overline{y}_{\mathbf{s}}$ non dipende dall'ordine in cui le unità si presentano nel campione, ma dipende dalle ripetizioni. Ogni y_i è sommata in (13.20) tante volte quante la corrispondente unità compare nel campione $\mathbf{s}$. Pertanto, $\overline{y}_{\mathbf{s}}$ è migliorabile usando il teorema di Rao-Blackwell. Per costruire la Rao-Blackwellizzazione di $\overline{y}_{\mathbf{s}}$, indichiamo con $\nu(\mathbf{s})$ il numero di unità distinte in

$\mathbf{s}$ (numero di unità in $r(\mathbf{s})$), e sia $n(i;\,\mathbf{s})$ il numero di volte in cui l'unità i compare nel campione $\mathbf{s}$. Vale l'ovvia relazione:

$$\sum_{i\in r(\mathbf{s})} n(i;\,\mathbf{s}) = n. \tag{13.21}$$

Inoltre, si può scrivere

$$\overline{y}_{\mathbf{s}} = \frac{1}{n}\sum_{i\in r(\mathbf{s})} y_i\, n(i;\,\mathbf{s}). \tag{13.22}$$

Per il calcolo della media condizionata di $\overline{y}_{\mathbf{s}}$ rispetto a $\boldsymbol{y}(r(\mathbf{s}))$, osserviamo in primo luogo che

$$\begin{aligned} E[\overline{y}_{\mathbf{s}}\,|\boldsymbol{y}(r(\mathbf{s}))] &= E\left[\frac{1}{n}\sum_{i\in r(\mathbf{s})} y_i\, n(i;\,\mathbf{s})\,\middle|\,\boldsymbol{y}(r(\mathbf{s}))\right] \\ &= \frac{1}{n}\sum_{i\in r(\mathbf{s})} E[y_i\, n(i;\,\mathbf{s})\,|\boldsymbol{y}(r(\mathbf{s}))] \\ &= \frac{1}{n}\sum_{i\in r(\mathbf{s})} y_i\, E[n(i;\,\mathbf{s})\,|r(\mathbf{s})]. \end{aligned} \tag{13.23}$$

Ragionando per simmetria, è poi facile vedere che i valori attesi $E[n(i;\,\mathbf{s})\,|r(\mathbf{s})]$ sono uguali per tutte le unità $i \in r(\mathbf{s})$. Indicando con $\overline{n}$ il loro valore comune, si ha cioè

$$E[n(i;\,\mathbf{s})\,|r(\mathbf{s})] = \overline{n} \;\; \textit{per ciascun } i \in r(\mathbf{s}). \tag{13.24}$$

Usando contemporaneamente (13.21) e (13.24) si ottiene quindi

$$n = \sum_{i\in r(\mathbf{s})} \overline{n} = \nu(\mathbf{s})\,\overline{n}$$

da cui segue che $\overline{n} = n/\nu(\mathbf{s})$, e quindi

$$E[n(i;\,\mathbf{s})\,|r(\mathbf{s})] = \frac{n}{\nu(\mathbf{s})} \;\; \textit{per ciascun } i \in r(\mathbf{s}). \tag{13.25}$$

Inserendo infine la (13.25) in (13.23) si ottiene:

$$\begin{aligned} E[\overline{y}_{\mathbf{s}}\,|\boldsymbol{y}(r(\mathbf{s}))] &= \frac{1}{n}\sum_{i\in r(\mathbf{s})} y_i\,\frac{n}{\nu(\mathbf{s})} \\ &= \frac{1}{\nu(\mathbf{s})}\sum_{i\in r(\mathbf{s})} y_i \\ &= \overline{y}_{r(\mathbf{s})} \;\; \textit{media campionaria per le unità distinte.} \end{aligned}$$

□

13.4 Non esistenza dello stimatore corretto di varianza minima

Nelle sezioni precedenti si è cercato di riportare l'inferenza da popolazioni finite nell'alveo generale della teoria dell'inferenza statistica, introducendo la statistica sufficiente minimale. Se il problema è quello di stimare un parametro $\theta = \theta(\boldsymbol{Y}_N)$, l'obiettivo ideale sarebbe quello di costruire uno stimatore "ottimo" secondo un qualche criterio.

Come già visto nel Capitolo 2, uno stimatore di minimo errore quadratico medio non esiste, e non solo nell'inferenza da popolazioni finite. È necessario quindi ripiegare su un obiettivo più modesto ma anche più ragionevole. La risposta "classica" della teoria della stima puntuale è quella di limitarsi ai soli stimatori corretti di θ, e di cercare tra essi, se esiste, quello di varianza uniformemente minima (*UMVUE = Uniformly Minimum Variance Unbiased Estimator*). Il risultato principale in questa direzione, nella statistica classica, è il *teorema di Lehmann-Scheffé*, il quale stabilisce condizioni (sufficienti) sotto cui esiste lo *UMVUE*.

Una statistica sufficiente T è *completa* se l'*unica* funzione $h(T)$ (a valori reali) tale che

$$E[h(T)] = 0 \;\; \textit{per ciascun } \boldsymbol{Y}_N \in \Omega_N \tag{13.26}$$

è quella identicamente nulla. In simboli:

$$E[h(T)] = 0 \;\; \textit{per ciascun } \boldsymbol{Y}_N \in \Omega_N \;\; \text{implica che} \;\; h(T) \equiv 0.$$

Non è difficile verificare che se una statistica sufficiente T è completa, allora è anche minimale.

Il *teorema di Lehmann-Scheffé* stabilisce che *se* (i) *esiste uno stimatore corretto di* θ *e* (ii) T *è una statistica sufficiente completa, allora esiste anche lo stimatore corretto di varianza uniformemente minima di* θ. Non è difficile verificare (Esercizio 13.7) che se T è sufficiente completa e se $\widehat{\theta}$ è un qualsiasi stimatore corretto di θ, allora lo stimatore corretto di varianza uniformemente minima di θ si può ottenere mediate Rao-Blackwellizzazione di $\widehat{\theta}$. In simboli:

$$\textit{Se } T \textit{ è sufficiente completa e } E[\widehat{\theta}] = \theta \;\; \text{allora} \;\; E[\widehat{\theta}\,|T] \textit{ è } UMVUE. \tag{13.27}$$

La Proposizione 13.4 contiene un risultato negativo, ovvero che la statistica sufficiente minimale $\boldsymbol{y}(r(\mathbf{s}))$ non è completa.

Proposizione 13.4. *La statistica sufficiente minimale* $\boldsymbol{y}(r(\mathbf{s}))$ *non è completa.*

Dimostrazione. Basta costruire una funzione $h(\boldsymbol{y}(r(\mathbf{s})))$ non identicamente nulla ma avente valore atteso pari a 0 qualunque sia il parametro $\boldsymbol{Y}_N$ della popolazione. Consideriamo a questo proposito la funzione:

$$\begin{aligned} h(\boldsymbol{y}(r(\mathbf{s}))) &= \begin{cases} \frac{1}{\pi_1} \; se \; 1 \in r(\mathbf{s}) \\ -\frac{1}{1-\pi_1} \; se \; 1 \notin r(\mathbf{s}) \end{cases} \\ &= \frac{1}{\pi_1}\delta(1;\, r(\mathbf{s})) - \frac{1}{1-\pi_1}(1 - \delta(1;\, r(\mathbf{s}))) \end{aligned}$$

la quale è non identicamente uguale a 0. Tuttavia, il suo valore atteso

$$\begin{aligned} E[h(\boldsymbol{y}(r(\mathbf{s})))] &= \frac{1}{\pi_1}E[\delta(1;\, r(\mathbf{s}))] - \frac{1}{1-\pi_1}(1 - E[\delta(1;\, r(\mathbf{s}))]) \\ &= \frac{1}{\pi_1}\,\pi_1 - \frac{1}{1-\pi_1}\,(1-\pi_1) \\ &= 0 \;\; per \; ciascun \; \boldsymbol{Y}_N \in \Omega_N \end{aligned}$$

è identicamente nullo. Tenendo infine conto che una statistica sufficiente completa, se esiste, è anche minimale, la dimostrazione è completata. □

Un'osservazione importante. La dimostrazione della Proposizione 13.4 sfrutta in modo decisivo il fatto che i dati campionari ridotti sono *modalità etichettate*, ovvero modalità che conservano l'informazione sull'unità a cui si riferiscono. Se le modalità campionarie venissero private delle etichette, ovvero se si "ricordassero" solo i valori y_i, $i \in \mathbf{s}$, e si "dimenticassero" le unità di riferimento, la dimostrazione della Proposizione 13.4 verrebbe meno.

L'esistenza di una statistica sufficiente completa è una condizione sufficiente ma non necessaria per l'esistenza di uno stimatore corretto di varianza uniformemente minima. Prima di enunciare e provare tale risultato è bene far riferimento al seguente esempio, che chiarisce l'elegante e semplice idea di base della dimostrazione (cfr. Basu (1971)).

Esempio 13.6. Sia $I_N = \{1, \ldots, N\}$ una popolazione finita di numerosità N, e si consideri un disegno ssr di numerosità n. Il problema è quello di stimare la media della popolazione, $\mu_y = \sum_{i=1}^N y_i/N$. Dato un arbitrario vettore $\mathbf{a}_N = [a_1, a_2, \ldots, a_N]$, sia $\mu_a = \sum_{i=1}^N a_i/N$. In particolare, se $\mathbf{a}_N = \boldsymbol{Y}_N$, cioè se $y_i = a_i$ per ciascuna unità i della popolazione, la media μ_y si riduce a μ_a. Si consideri poi, con ovvia simbologia, lo stimatore di μ_y

$$\begin{aligned} t_a &= \frac{1}{n}\sum_{i\in\mathbf{s}} y_i - \frac{1}{n}\sum_{i\in\mathbf{s}} a_i + \mu_a \\ &= \overline{y}_{\mathbf{s}} - \overline{a}_{\mathbf{s}} + \mu_a \end{aligned} \tag{13.28}$$

in cui $\overline{y}_{\mathbf{s}}$ è la media campionaria delle y_i, e $\overline{a}_{\mathbf{s}}$ è la media campionaria delle a_i.

Se il disegno è ssr si ha $E[\overline{y}_{\mathbf{s}}] = \mu_y$, $E[\overline{a}_{\mathbf{s}}] = \mu_a$, per cui è anche $E[t_a] = \mu_y$ comunque si scelga il vettore $\mathbf{a}_N$. Inoltre, se il parametro della popolazione,

$\boldsymbol{Y}_N$, coincide con $\mathbf{a}_N$ si ha $\mu_y = \mu_a$ e $t_a = \mu_a$, per cui, in questo caso speciale, lo stimatore t_a è identicamente uguale al parametro da stimare. La sua varianza, di conseguenza, è nulla. In simboli:

$$E[t_a] = \mu_y \quad \textit{qualunque siano } \boldsymbol{Y}_N \in I\!R^N \textit{ e } \mathbf{a}_N \in I\!R^N;$$
$$\textit{se } \boldsymbol{Y}_N = \mathbf{a}_N \text{ allora } \mu_y = \mu_a,\ t_a \equiv \mu_y,\ V(t_a) = 0.$$

Indichiamo infine con $\mathcal{D}_u$ la classe di *tutti* gli stimatori (corretti) del tipo (13.28), al variare di $\mathbf{a}_N$ in $I\!R^N$. In simboli:

$$\mathcal{D}_u = \{t_a = \overline{y}_{\mathbf{s}} - \overline{a}_{\mathbf{s}} + \mu_a;\ \ \mathbf{a}_N \in I\!R^N\}.$$

Se esistesse uno stimatore corretto t^* di μ_y di varianza uniformemente minima, esso dovrebbe avere varianza più piccola di un qualunque stimatore t_a della classe $\mathcal{D}_u$. Dovrebbe in altre parole potersi scrivere

$$V(t^*) \leqslant V(t_a) \quad \textit{qualunque siano } \boldsymbol{Y}_N \in I\!R^N \textit{ e } \mathbf{a}_N \in I\!R^N. \tag{13.29}$$

Ma per $\boldsymbol{Y}_N = \mathbf{a}_N$ si ha $V(t_a) = 0$, e quindi dalla (13.29) si trae che

$$V(t^*) = 0 \ \textit{ se } \boldsymbol{Y}_N = \mathbf{a}_N, \ \ \textit{qualunque sia } \mathbf{a}_N \in I\!R^N$$

il che equivale a scrivere

$$V(t^*) = 0 \ \textit{ qualunque sia } \boldsymbol{Y}_N \in I\!R^N. \tag{13.30}$$

Chiaramente, la (13.30) può aver luogo solo se t^* coincide sempre con la media μ_y della popolazione, qualunque sia il campione $\mathbf{s}$, il che è impossibile. La conclusione è quindi che se il disegno è ssr non esiste lo stimatore corretto di varianza uniformemente minima di μ_y. □

Proposizione 13.5. *Sia $I_N = \{1, \ldots, N\}$ una popolazione finita di numerosità N, da cui si seleziona un campione mediante un disegno $(\mathcal{S},\ p(\cdot))$. Detto $\theta = \theta(\boldsymbol{Y}_N)$ il parametro di interesse, non esiste lo stimatore corretto di varianza uniformemente minima di θ.*

Dimostrazione. La dimostrazione usa le stesse idee dell'Esempio 13.6. Sia $\widehat{\theta}(\boldsymbol{y}(\mathbf{s}))$ uno stimatore corretto di θ. Dato un arbitrario vettore $\mathbf{a}_N = [a_1, a_2, \ldots, a_N]$, sia $\theta(\mathbf{a}_N)$ il valore assunto da θ quando $\boldsymbol{Y}_N = \mathbf{a}_N$. Poniamo inoltre $\mathbf{a}(\mathbf{s}) = \{(i, a_i) \ \ i \in \mathbf{s}\}$ l'insieme dei valori etichettati a_i corrispondenti alle unità del campione $\mathbf{s}$.

Si definisca poi lo stimatore di θ:

$$\widehat{\theta}_a = \widehat{\theta}(\boldsymbol{y}(\mathbf{s})) - \widehat{\theta}(\mathbf{a}(\mathbf{s})) + \theta(\mathbf{a}_N). \tag{13.31}$$

Essendo $\widehat{\theta}$ corretto rispetto al disegno $(\mathcal{S},\ p(\cdot))$, si ha $E[\widehat{\theta}(\boldsymbol{y}(\mathbf{s}))] = \theta(\boldsymbol{Y}_N)$, $E[\widehat{\theta}(\mathbf{a}(\mathbf{s}))] = \theta(\mathbf{a}_N)$, per cui è anche $E[t_a] = \theta(\boldsymbol{Y}_N)$ qualunque sia il vettore $\mathbf{a}_N$. Inoltre, se il parametro della popolazione, $\boldsymbol{Y}_N$, coincide con $\mathbf{a}_N$ si ha

$\theta(\boldsymbol{Y}_N) = \theta(\mathbf{a}_N)$ e $\widehat{\theta}_a \equiv \theta(\mathbf{a}_N)$. Ne consegue che se $\boldsymbol{Y}_N = \mathbf{a}_N$, la varianza di $\widehat{\theta}_a$ è pari a 0. Si può scrivere, in sintesi,

$$E[\widehat{\theta}_a] = \theta(\boldsymbol{Y}_N) \quad \textit{qualunque siano } \boldsymbol{Y}_N \in \Omega^N \textit{ e } \mathbf{a}_N \in \Omega^N;$$
$$\textit{se } \boldsymbol{Y}_N = \mathbf{a}_N \text{ allora } \theta(\boldsymbol{Y}_N) = \theta(\mathbf{a}_N),\ \widehat{\theta}_a \equiv \theta(\boldsymbol{Y}_N),\ V(\widehat{\theta}_a) = 0.$$

Indichiamo infine con $\mathcal{D}_u$ la classe di *tutti* gli stimatori (corretti) del tipo (13.31), al variare di $\mathbf{a}_N$ in $I\!R^N$. In simboli:

$$\mathcal{D}_u = \{\widehat{\theta}_a = \widehat{\theta}(\boldsymbol{y}(\mathbf{s})) - \widehat{\theta}(\mathbf{a}(\mathbf{s})) + \theta(\mathbf{a}_N);\ \ \mathbf{a}_N \in I\!R^N\}.$$

Se esistesse uno stimatore corretto $\widehat{\theta}^*$ di θ di varianza uniformemente minima, esso dovrebbe avere varianza più piccola di un qualunque stimatore $\widehat{\theta}_a$ della classe $\mathcal{D}_u$:

$$V(\widehat{\theta}^*) \leqslant V(\widehat{\theta}_a) \quad \textit{qualunque siano } \boldsymbol{Y}_N \in \Omega^N \textit{ e } \mathbf{a}_N \in \Omega^N. \tag{13.32}$$

Ma se $\boldsymbol{Y}_N = \mathbf{a}_N$ si ha $V(\widehat{\theta}_a) = 0$, per cui dalla (13.32) discende che

$$V(\widehat{\theta}^*) = 0 \ \ \textit{se } \boldsymbol{Y}_N = \mathbf{a}_N, \ \ \textit{qualunque sia } \mathbf{a}_N \in \Omega^N$$

che equivale a

$$V(\widehat{\theta}^*) = 0 \ \ \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega^N. \tag{13.33}$$

La (13.33) può verificarsi soltanto se $\widehat{\theta}^*$ coincide sempre con $\theta(\boldsymbol{Y}_N)$, qualunque siano il campione $\mathbf{s}$ e $\boldsymbol{Y}_N$ in Ω_N. Ma questo è impossibile, per cui non esiste lo stimatore corretto di varianza uniformemente minima di $\theta(\boldsymbol{Y}_N)$. □

Esattamente come la Proposizione 13.4, anche la Proposizione 13.5 si basa sull'assunzione che i dati campionari consistano di modalità etichettate. Quest'ipotesi è fondamentale per costruire l'insieme di valori etichettati $\mathbf{a}(\mathbf{s})$. Se cadesse l'ipotesi che i dati campionari consistano di modalità etichettate, cadrebbe anche la dimostrazione della Proposizione 13.5. Per ulteriori risultati sull'esistenza di stimatori corretti di varianza uniformemente minima in casi speciali si rinvia al volume di Cassel *e altri* (1977).

13.5 La nozione di ammissibilità di stimatori e strategie

La non esistenza di uno stimatore ottimo nella classe degli stimatori corretti (e neanche in classi più ristrette: Cassel *e altri* (1977)) sposta l'interesse verso la ricerca di stimatori con proprietà "ragionevolmente buone". In questa direzione un requisito minimale, ma intuitivamente rilevante, che dovrebbe possedere uno stimatore è quello di non essere peggiore di nessun altro stimatore, perlomeno all'interno di una determinata classe. Questo porta all'introduzione della nozione di ammissibilità.

Data una popolazione finita di N unità, si supponga di selezionare da essa un campione mediante un disegno $(\mathcal{S},\ p(\cdot))$. Per il momento assumeremo tale disegno *fissato*. Sia $\mathcal{T}$ una classe di stimatori di un parametro $\theta = \theta(\boldsymbol{Y}_N)$.

1. Uno stimatore $\widehat{\theta}_0$ in $\mathcal{T}$ è *non migliore* di uno stimatore $\widehat{\theta}_1$ in $\mathcal{T}$ (equivalentemente, $\widehat{\theta}_1$ è *non peggiore* di $\widehat{\theta}_0$) se

$$MSE(\widehat{\theta}_1) \leqslant MSE(\widehat{\theta}_0) \quad \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N. \tag{13.34}$$

2. Uno stimatore $\widehat{\theta}_0$ in $\mathcal{T}$ è *peggiore* di uno stimatore $\widehat{\theta}_1$ in $\mathcal{T}$ (equivalentemente, $\widehat{\theta}_1$ è *migliore* di $\widehat{\theta}_0$, o anche $\widehat{\theta}_1$ *domina* $\widehat{\theta}_0$) se

$$MSE(\widehat{\theta}_1) \leqslant MSE(\widehat{\theta}_0) \quad \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N; \tag{13.35}$$

$$MSE(\widehat{\theta}_1) < MSE(\widehat{\theta}_0) \quad \textit{per almeno un } \boldsymbol{Y}_N \in \Omega_N. \tag{13.36}$$

Gli errori quadratici medi in (13.34)-(13.36) sono calcolati, come detto, rispetto ad un prefissato disegno di campionamento.

Quando si sceglie uno stimatore di θ in una classe $\mathcal{T}$ di stimatori,un requisito molto naturale, e per molti aspetti minimale, è che nella classe $\mathcal{T}$ non vi sia nessuno stimatore migliore di quello scelto. Questo porta alla nozione di ammissibilità. Uno stimatore $\widehat{\theta}_0$ in $\mathcal{T}$ è *ammissibile* nella classe $\mathcal{T}$ di stimatori (sempre rispetto ad un fissato disegno campionario) se non esiste in $\mathcal{T}$ nessuno stimatore migliore di $\widehat{\theta}_0$.

L'ammissibilità di uno stimatore è effettivamente un requisito molto debole per uno stimatore, ma tuttavia importante. Una conseguenza immediata del teorema di Rao-Blackwell è che *tutti gli stimatori che non dipendono dalla statistica sufficiente minimale* $\boldsymbol{y}(r(\mathbf{s}))$ *non sono ammissibili.* Il teorema di Rao-Blackwell va anche un passo in avanti, in quanto mostra che se $\widehat{\theta}$ non è funzione di $\boldsymbol{y}(r(\mathbf{s}))$, allora è *peggiore* di $E[\widehat{\theta}\,|\boldsymbol{y}(r(\mathbf{s}))]$. In altre parole, il teorema di Rao-Blackwell insegna a riconoscere stimatori che non sono ammissibili, anche se nulla dice sull'ammissibilità o meno di $E[\widehat{\theta}\,|\boldsymbol{y}(r(\mathbf{s}))]$.

Nelle definizioni precedenti, come più volte sottolineato, il disegno campionario $(\mathcal{S},\ p(\cdot))$ è assunto fissato. Tuttavia, lo statistico sceglie la *coppia (stimatore, disegno campionario)* ossia la *strategia* di campionamento $((\mathcal{S},\ p(\cdot)),\ \widehat{\theta})$. Le nozioni dianzi introdotte possono facilmente estendersi a strategie. Nel seguito si indicherà con $\mathcal{ST}$ una classe di strategie di campionamento.

1. La strategia $((\mathcal{S}_0, p_0(\cdot)), \widehat{\theta}_0)$ in $\mathcal{ST}$ è *non migliore* della strategia $((\mathcal{S}_1, p_1(\cdot)),$ $\widehat{\theta}_1)$ in $\mathcal{ST}$ (equivalentemente, $((\mathcal{S}_1, p_1(\cdot)), \widehat{\theta}_1)$ è *non peggiore* di $((\mathcal{S}_0, p_1 0(\cdot)),$ $\widehat{\theta}_0))$ se

$$MSE_1(\widehat{\theta}_1) \leqslant MSE_0(\widehat{\theta}_0) \quad \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N\,, \tag{13.37}$$

dove $MSE_0(\widehat{\theta}_0)$ $(MSE_1(\widehat{\theta}_1))$ indica l'errore quadratico medio di $\widehat{\theta}_0$ $(\widehat{\theta}_1)$ calcolato rispetto al disegno $(\mathcal{S}_0,\ p_0(\cdot))$ $((\mathcal{S}_1,\ p_1(\cdot)))$.

2. La strategia $((\mathcal{S}_0, p_0(\cdot)), \widehat{\theta}_0)$ in $\mathcal{ST}$ è *peggiore* della strategia $((\mathcal{S}_1, p_1(\cdot)),\ \widehat{\theta}_1)$ in $\mathcal{ST}$ (equivalentemente, $((\mathcal{S}_1,\ p_1(\cdot)),\ \widehat{\theta}_1)$ è *migliore* di $((\mathcal{S}_0,\ p_0(\cdot)),\ \widehat{\theta}_0)$, o anche $((\mathcal{S}_1,\ p_1(\cdot)),\ \widehat{\theta}_1)$ *domina* $((\mathcal{S}_0,\ p_0(\cdot)),\ \widehat{\theta}_0)$) se

$$MSE_1(\widehat{\theta}_1) \leqslant MSE_0(\widehat{\theta}_0) \quad \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N; \tag{13.38}$$

$$MSE_1(\widehat{\theta}_1) < MSE_0(\widehat{\theta}_0) \quad \textit{per almeno un } \boldsymbol{Y}_N \in \Omega_N. \tag{13.39}$$

La nozione di ammissibilità di stimatori si estende facilmente a strategie. Una strategia $((\mathcal{S}_0, p_0(\cdot)),\ \widehat{\theta}_0)$ in $\mathcal{ST}$ è *ammissibile* nella classe $\mathcal{ST}$ di strategie se non esiste in $\mathcal{ST}$ nessuna strategia migliore di $((\mathcal{S}_0,\ p_0(\cdot)),\ \widehat{\theta}_0)$.

13.6 La tecnica di contrazione di stimatori

La tecnica di contrazione (*shrinkage*) di un stimatore è stata introdotta da Stein (1956) come metodo per migliorare il vettore delle medie campionarie per la stima del vettore dei valori attesi di una distribuzione multinormale di dimensione maggiore di 2, ed ha ricevuto da allora notevole attenzione in statistica. Qui tale tecnica verrà brevemente presentata; nel capitolo successivo verrà applicata al miglioramento dello stimatore di Horvitz-Thompson.

Dato un disegno campionario $(\mathcal{S},\ p(\cdot))$, sia $\widehat{\theta}$ uno stimatore corretto di un parametro $\theta = \theta(\boldsymbol{Y}_N)$ di interesse: $E[\widehat{\theta}] = \theta$. Detta $V(\widehat{\theta})$ la varianza di $\widehat{\theta}$, si consideri come stimatore alternativo a $\widehat{\theta}$ lo stimatore definito nel seguente modo:

$$\widehat{\theta}_{sh} = c\,\widehat{\theta} \tag{13.40}$$

dove c è un numero reale. Lo stimatore (13.40) è distorto con errore quadratico medio pari a

$$\begin{aligned} MSE(\widehat{\theta}_{sh}) &= V(\widehat{\theta}) + (E[\widehat{\theta}_{sh}] - \theta)^2 \\ &= c^2\, V(\widehat{\theta}) + (c-1)^2\theta^2\,. \end{aligned} \tag{13.41}$$

Chiaramente, lo stimatore $\widehat{\theta}_{sh}$ rappresenta un miglioramento dello stimatore iniziale $\widehat{\theta}$ per quei valori di c che rendono negativa la differenza tra i due errori quadratici medi:

$$f(c) = MSE(\widehat{\theta}_{sh}) - MSE(\widehat{\theta}) < 0\,. \tag{13.42}$$

I valori di c che soddisfano la (13.42) sono interni all'intervallo che ha come estremi le due soluzioni dell'equazione in c di secondo grado $f(c) = 0$. Formalmente:

$$\frac{\theta^2 - V(\widehat{\theta})}{\theta^2 + V(\widehat{\theta})} < c < 1\,.$$

Lo stimatore $\widehat{\theta}_{sh} = c\,\widehat{\theta}$, con $|c| < 1$, è detto "stimatore per contrazione" (*shrinkage*) di θ. Il valore ottimo del numero c si ricava minimizzando l'errore quadratico medio di $\widehat{\theta}_{sh}$. A tale scopo, uguagliando a zero la derivata rispetto a c della (13.41), si ottiene l'equazione

$$\frac{\partial MSE(\widehat{\theta}_{sh})}{\partial \theta} = 2\,c\,V(\widehat{\theta}) + 2\,(c-1)\,\theta^2 = 0 \tag{13.43}$$

che fornisce il valore

$$c_{opt} = \frac{\theta^2}{\theta^2 + V(\widehat{\theta})} = \frac{\theta^2}{E[\widehat{\theta}^2]}\,. \tag{13.44}$$

Come detto, ogni valore $c_{opt} \leqslant c < 1$ è tale che $MSE(\widehat{\theta}_{sh}) < V(\widehat{\theta})$. Il vantaggio massimo, ovviamente, si ottiene prendendo $c = c_{opt}$. Vi è però un problema. I valori di θ e di $E[\widehat{\theta}^2]$ dipendono dall'intero parametro $\boldsymbol{Y}_N$ della popolazione, che è incognito. Pertanto, lo stesso valore di c_{opt} in (13.44) dipende da $\boldsymbol{Y}_N$, e quindi non è calcolabile. In simboli:

$$c_{opt} = c_{opt}(\boldsymbol{Y}_N).$$

L'idea è allora quella di prendere il massimo valore di $c_{opt}(\boldsymbol{Y}_N)$, rispetto a tutti i possibili valori del parametro $\boldsymbol{Y}_N$ della popolazione. In altre parole, l'idea è quella di calcolare:

$$\begin{aligned} c^* &= \max_{\mathbf{Y}_N \in \Omega_N} c_{opt}(\boldsymbol{Y}_N) \\ &= \max_{\mathbf{Y}_N \in \Omega_N} \frac{\theta^2}{E[\widehat{\theta}^2]} \\ &= \frac{1}{\min_{\mathbf{Y}_N \in \Omega_N} \frac{E[\widehat{\theta}^2]}{\theta^2}} \qquad (13.45) \\ &= \frac{1}{1 + \min_{\mathbf{Y}_N \in \Omega_N} \frac{V(\widehat{\theta})}{\theta^2}}. \qquad (13.46) \end{aligned}$$

Quanto detto è illustrato in Fig. 13.3.

Chiaramente, ogni valore $c^* \leqslant c < 1$ è tale che l'errore quadratico medio di $\widehat{\theta}_{sh} = c\,\widehat{\theta}$ è più piccolo di quello di $\widehat{\theta}$. Ancora una volta, il vantaggio massimo si ottiene prendendo $c = c^*$. Ad esso corrisponde lo stimatore:

$$\widehat{\theta}_{hs*} = c^*\,\widehat{\theta}. \tag{13.47}$$

Per quanto riguarda il valore di c^*, due sono le possibilità:

- $c^* = 1$: in questo caso lo stimatore (13.47) coincide con $\widehat{\theta}$, e la tecnica di contrazione non dà nessun miglioramento;
- $c^* < 1$: in questo caso lo stimatore (13.47) è migliore di $\widehat{\theta}$, nel senso che possiede un errore quadratico medio più piccolo.

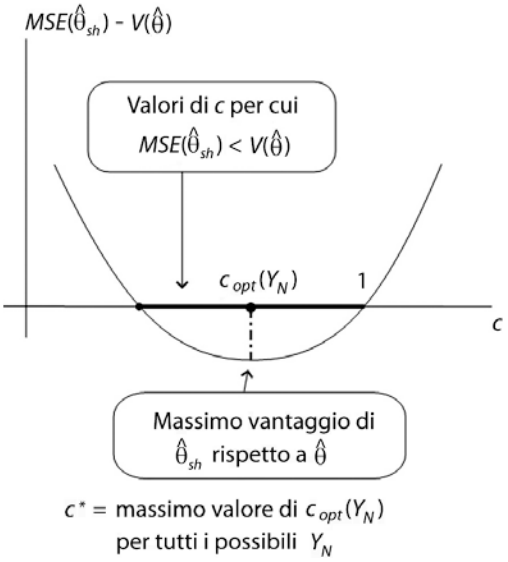


Fig. 13.3 Contrazione dello stimatore $\widehat{\theta}$

Un'analisi un po' più dettagliata della relazione (13.46) consente di capire meglio quando si verifica il caso $c^* = 1$. Infatti, è evidente che $c^* = 1$ se e solo se per ciascun possibile valore del parametro di interesse θ esiste uno "speciale" $\widetilde{\boldsymbol{Y}}_N$ tale che $\theta(\widetilde{\boldsymbol{Y}}_N) = \theta$ e $V(\widehat{\theta}) = 0$. Si osservi che quest'ultima uguaglianza, essendo $\widehat{\theta}$ corretto, equivale (con ovvia simbologia) a $\widehat{\theta}(\widetilde{\boldsymbol{y}}(\mathbf{s})) = \theta(\widetilde{\boldsymbol{Y}}_N)$ $(= \theta)$ qualunque sia il campione $\mathbf{s}$.

Esempio 13.7. Si consideri una popolazione finita di $N = 3$ unità, di cui si deve stimare la media $\mu_y = (y_1 + y_2 + y_3)/3$. Il disegno campionario è definito come segue:

$$\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4, \mathbf{s}_5, \mathbf{s}_6\};$$
$$\mathbf{s}_1 = \{1\},\ \mathbf{s}_2 = \{2\},\ \mathbf{s}_3 = \{3\},\ \mathbf{s}_4 = \{1, 2\},\ \mathbf{s}_5 = \{1, 3\},\ \mathbf{s}_6 = \{2, 3\};$$
$$p(\mathbf{s}_1) = 1/9,\ p(\mathbf{s}_2) = 1/9,\ p(\mathbf{s}_3) = 1/9,\ p(\mathbf{s}_4) = 2/9,\ p(\mathbf{s}_5) = 2/9,$$
$$p(\mathbf{s}_6) = 2/9.$$

Per stimare la media μ_y, si consideri pi lo stimatore

$$\widehat{\mu} = \frac{3}{5} \sum_{i \in \mathbf{s}} y_i. \tag{13.48}$$

Tabella 13.1 Valori dello stimatore $\widehat{\mu}$ per i campioni $\mathbf{s}_1, \ldots, \mathbf{s}_6$

Campione	*Probabilità*	$\widehat{\mu}$
$\{1\}$	1/9	$\frac{3}{5}y_1$
$\{2\}$	1/9	$\frac{3}{5}y_2$
$\{3\}$	1/9	$\frac{3}{5}y_3$
$\{1, 2\}$	2/9	$\frac{3}{5}(y_1 + y_2)$
$\{1, 3\}$	2/9	$\frac{3}{5}(y_1 + y_3)$
$\{2, 3\}$	2/9	$\frac{3}{5}(y_2 + y_3)$

È immediato verificare che lo stimatore $\widehat{\mu}$ è corretto. Con pochi facili calcoli, inoltre, si ha che:

$$\begin{aligned} E[\widehat{\mu}^2] &= \frac{1}{9} \times \frac{9}{25} \times (y_1^2 + y_2^2 + y_3^2) + \frac{2}{9} \times \frac{9}{25} \\ &\quad \times (2y_1^2 + 2y_2^2 + 2y_3^2 + 2y_1y_2 + 2y_1y_3 + 2y_2y_3) \\ &= \frac{1}{9}\left\{\frac{27}{25}(y_1^2 + y_2^2 + y_3^2) + \frac{18}{25}(y_1 + y_2 + y_3)^2\right\} \\ &= \frac{27}{9 \times 25}(y_1^2 + y_2^2 + y_3^2) + \frac{18}{25}\mu_y^2 \end{aligned}$$

da cui si ottiene

$$\frac{E[\widehat{\mu}^2]}{\mu_y^2} = \frac{27}{25}\frac{y_1^2 + y_2^2 + y_3^2}{(y_1 + y_2 + y_3)^2} + \frac{18}{25}.$$

In base alla (13.45), bisogna ora minimizzare il termine

$$\begin{aligned} g(y_1, y_2, y_3) &= \frac{E[\widehat{\mu}^2]}{\mu_y^2} \\ &= \frac{27}{25}\frac{y_1^2 + y_2^2 + y_3^2}{(y_1 + y_2 + y_3)^2} + \frac{18}{25}. \end{aligned} \tag{13.49}$$

Derivando la (13.49) rispetto a y_1, y_2, y_3 e annullando tali derivate, si ha

$$\begin{aligned} \frac{\partial g}{\partial y_1} &= \frac{\frac{54}{25}\left\{y_1(y_1 + y_2 + y_3)^2 - (y_1 + y_2 + y_3)(y_1^2 + y_2^2 + y_3^2)\right\}}{(y_1 + y_2 + y_3)^4} = 0 \\ \frac{\partial g}{\partial y_2} &= \frac{\frac{54}{25}\left\{y_2(y_1 + y_2 + y_3)^2 - (y_1 + y_2 + y_3)(y_1^2 + y_2^2 + y_3^2)\right\}}{(y_1 + y_2 + y_3)^4} = 0 \\ \frac{\partial g}{\partial y_3} &= \frac{\frac{54}{25}\left\{y_3(y_1 + y_2 + y_3)^2 - (y_1 + y_2 + y_3)(y_1^2 + y_2^2 + y_3^2)\right\}}{(y_1 + y_2 + y_3)^4} = 0 \end{aligned} \tag{13.50}$$

e l'unica soluzione delle (13.50) è, come facilmente si vede,

$$y_1 = y_2 = y_3 = \mu_y. \tag{13.51}$$

Dalla (13.51) si ricava che

$$\begin{aligned} \min \frac{E[\widehat{\mu}^2]}{\mu_y^2} &= \frac{27}{25}\frac{3}{9}\frac{\mu_y^2}{\mu_y^2} + \frac{18}{25} \\ &= \frac{27}{25} \end{aligned}$$

da cui si ottiene, in base alla (13.45)

$$c^* = \frac{25}{27}.$$

Usando infine la (13.47), si conclude che lo stimatore

$$\widehat{\mu}_{hs*} = \frac{25}{27}\,\widehat{\mu}$$

ha errore quadratico medio più piccolo di $\widehat{\mu}$. □

Esempio 13.8. Si consideri una popolazione finita di $N = 3$ unità, di cui si deve stimare la media $\mu_y = (y_1 + y_2 + y_3)/3$. Il disegno campionario è di seguito specificato:

$$\begin{gathered} \mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2\}; \\ \mathbf{s}_1 = \{1\},\ \mathbf{s}_2 = \{2, 3\}; \\ p(\mathbf{s}_1) = 1/2,\ p(\mathbf{s}_2) = 1/2. \end{gathered}$$

Per stimare la media μ_y, si consideri poi lo stimatore

$$\begin{aligned} \widehat{\mu} &= \frac{2}{3}\sum_{i\in\mathbf{s}} y_i \\ &= \begin{cases} \frac{2}{3}\,y_1 \ \textit{se}\ \mathbf{s} = \mathbf{s}_1 \\ \frac{2}{3}\,(y_2 + y_3) \ \textit{se}\ \mathbf{s} = \mathbf{s}_1\,. \end{cases} \end{aligned} \tag{13.52}$$

Si vede subito che lo stimatore (13.52) è corretto. Inoltre, si ha

$$\begin{aligned} E[\widehat{\mu}^2] &= \frac{1}{2} \times \frac{4}{9} \times y_1^2 + \frac{1}{2} \times \frac{4}{9} \times (y_2 + y_3)^2 \\ &= \frac{2}{9}\left\{y_1^2 + (y_2 + y_3)^3\right\}. \end{aligned}$$

Il rapporto $E[\widehat{\mu}^2]/\mu_y^2$ raggiunge il suo valore minimo, pari a 1, per

$$y_1 = \tfrac{3}{2}\,\mu_y,\ y_2 = \tfrac{3}{4}\,\mu_y,\ y_3 = \tfrac{3}{4}\,\mu_y$$

così che

$$c^* = \frac{1}{\min(E[\widehat{\mu}^2]/\mu_y^2)} = 1.$$

Pertanto, usando la (13.47) si ha che lo stimatore $\widehat{\mu}_{hs*} = c^*\widehat{\mu}$ coincide con $\widehat{\mu}$. La tecnica di contrazione, in questo caso, non riesce a migliorare lo stimatore $\widehat{\mu}$. □

Esempio 13.9. Consideriamo un disegno semplice senza ripetizione di ampiezza n e supponiamo di voler stimare la media della popolazione. L'applicazione della tecnica di contrazione alla media campionaria $\overline{y}_{\mathbf{s}}$ porta alla minimizzazione della seguente quantità

$$\begin{aligned}\frac{E[\overline{y}_{\mathbf{s}}^2]}{\mu_y^2} &= \frac{V(\overline{y}_{\mathbf{s}}) + \mu_y^2}{\mu_y^2} \\ &= 1 + \frac{V(\overline{y}_{\mathbf{s}})}{\mu_y^2} \\ &= 1 + \frac{\left(\frac{1}{n} - \frac{1}{N}\right) S_y^2}{\mu_y^2}\end{aligned} \tag{13.53}$$

dove S_y^2 è la varianza corretta della popolazione. Ora, qualunque sia μ_y, se $y_1 = y_2 = \cdots = y_N = \mu_y$ si ha $S_y^2 = 0$. Quindi, il valore minimo della (13.53) è

$$\min_{\mathbf{Y}_N \in \mathbb{R}^N} \frac{E[\overline{y}_{\mathbf{s}}^2]}{\mu_y^2} = 1.$$

Ne consegue, per la (13.45), che $c^* = 1$. La tecnica di contrazione, in questo caso, non riesce a migliorare la media campionaria. □

Esempio 13.10. Consideriamo un disegno semplice con ripetizione di ampiezza n e supponiamo di voler stimare la media μ_y della popolazione. Come stimatore "iniziale" di μ_y consideriamo la media campionaria delle unità distinte introdotta nell'Esempio 13.5:

$$\overline{y}_{r(\mathbf{s})} = \frac{1}{\nu(\mathbf{s})} \sum_{i \in r(\mathbf{s})} y_i.$$

Come visto, $\overline{y}_{r(\mathbf{s})}$ è un stimatore corretto di μ_y.

L'applicazione della tecnica di contrazione a $\overline{y}_{r(\mathbf{s})}$ porta alla minimizzazione della seguente quantità:

$$\begin{aligned}\frac{E[\overline{y}_{r(\mathbf{s})}^2]}{\mu_y^2} &= \frac{V(\overline{y}_{r(\mathbf{s})}) + \mu_y^2}{\mu_y^2} \\ &= 1 + \frac{V(\overline{y}_{r(\mathbf{s})})}{\mu_y^2}.\end{aligned} \tag{13.54}$$

Qualunque sia la media μ_y della popolazione, se $y_1 = y_2 = \cdots = y_N = \mu_y$ si ha $\overline{y}_{r(\mathbf{s})} = \mu_y$, e quindi $V(\overline{y}_{\mathbf{s}}) = 0$. Ne consegue che il valore minimo della (13.54) è 1:

$$\min_{\mathbf{Y}_N \in I\!R^N} \frac{E[\overline{y}^2_{r(\mathbf{s})}]}{\mu_y^2} = 1$$

e che, per la (13.45), deve essere $c^* = 1$. La tecnica di contrazione non riesce dunque a migliorare la media campionaria delle unità distinte. □

Esercizi

13.1. Provare che se una statistica U induce una partizione più fine di quella indotta dalla statistica T, allora è possibile esprimere T come funzione di U.

Suggerimento. Se la partizione indotta da U è più fine di quella indotta da T, allora T assume lo stesso valore per tutti gli $y(\mathbf{s}) \in U^{-1}(u)$, per ciascun $u \in \mathcal{U}$. Detto t_u tale valore, basta porre $f(u) = t_u$, $u \in \mathcal{U}$, e osservare che $T = f(U)$.

13.2. Con riferimento all'Esempio 13.2, provare che la statistica $V = v(\boldsymbol{y}(\mathbf{s}))$ definita da

$$v(\boldsymbol{y}(\mathbf{s}_1)) = v(\boldsymbol{y}(\mathbf{s}_2)) = v(\boldsymbol{y}(\mathbf{s}_4)) = 0; \;\; v(\boldsymbol{y}(\mathbf{s}_3)) = v(\boldsymbol{y}(\mathbf{s}_6)) = 1; \;\; v(\boldsymbol{y}(\mathbf{s}_5)) = 2$$

è sufficiente minimale.

13.3. Data una popolazione finita $I_N = \{1, 2, \ldots, N\}$ di N unità, si supponga fissato *a priori* per ogni unità un numero $p_i > 0$, con $p_1 + p_2 + \cdots + p_N = 1$. Si consideri un disegno campionario (*ppswor*) $(\mathcal{S},\ p(\cdot))$, in cui $\mathcal{S}$ è l'insieme delle disposizioni senza ripetizione di classe 2 delle unità della popolazione; ogni campione $\mathbf{s} \in \mathcal{S}$ è quindi rappresentabile come coppia ordinata (i_1, i_2), dove i_1 e i_2 sono rispettivamente la prima e la seconda unità del campione $(i_1 \neq i_2)$. Il campione $\mathbf{s} = (i_1, i_2)$ ha probabilità $p(\mathbf{s}) = p_{i_1} p_{i_2}/(1 - p_{i_1})$. In sostanza, la prima unità del campione è selezionata con probabilità p_{i_1} e la seconda, data la prima, con probabilità $p_{i_2}/(1 - p_{i_1})$.

a. Verificare che $t_1 = y_{i_1}/(N p_{i_1})$ è uno stimatore corretto della media della popolazione.
b. Se $r(\mathbf{s}) = \{i, j\}$, mostrare che

$$E[t_1 \,|\boldsymbol{y}(r(\mathbf{s}))] = \frac{1}{N}\left\{\frac{y_i}{p_i}\frac{1 - p_j}{2 - p_i - p_j} + \frac{y_j}{p_j}\frac{1 - p_i}{2 - p_i - p_j}\right\}.$$

13.4. Con riferimento all'Esercizio 13.3, si ponga

$$t_2 = \frac{1}{N}\left(y_{i_1} + y_{i_2}\frac{1 - p_{i_1}}{p_{i_2}}\right).$$

a. Verificare che $t = (t_1 + t_2)/2$ è uno stimatore corretto della media della popolazione.
b. Se $r(\mathbf{s}) = \{i, j\}$, mostrare che

$$E[t\,|\boldsymbol{y}(r(\mathbf{s}))] = \frac{1}{N}\left\{\frac{y_i}{p_i}\frac{1-p_j}{2-p_i-p_j} + \frac{y_j}{p_j}\frac{1-p_i}{2-p_i-p_j}\right\}.$$

13.5. Data una popolazione finita $I_N = \{1, 2, \ldots, N\}$ di N unità, si consideri il disegno campionario $(\mathcal{S}, p(\cdot))$ qui sotto specificato.

– $\mathcal{S}$ è l'insieme delle disposizioni senza ripetizioni di classe 3 delle unità della popolazione; ogni campione $\mathbf{s} \in \mathcal{S}$ è rappresentabile come terna ordinata (i_1, i_2, i_3), dove i_1, i_2, i_3 sono rispettivamente la prima, la seconda, la terza unità del campione $(i_1 \neq i_2 \neq i_3)$;
– il generico campione $\mathbf{s} = (i_1, i_2, i_3)$ ha probabilità $p(\mathbf{s}) = \frac{1}{N(N-1)(N-2)}$.
a. Provare che

$$t = 0.2y_{i_1} + 0.5y_{i_2} + 0.2y_{i_3}$$

è uno stimatore corretto della media della popolazione.
b. Provare che la Rao-Blackwellizzazione di t è lo stimatore media campionaria.

13.6. Data una popolazione finita $I_N = \{1, 2, \ldots, N\}$ di N unità, si supponga assegnato *a priori* per ogni unità un numero $p_i > 0$, con $p_1 + p_2 + \cdots + p_N = 1$. Si consideri il disegno campionario (*ppswr*) $(\mathcal{S}, p(\cdot))$, qui sotto specificato.

– $\mathcal{S} = I_N \times I_N \times I_N$ è l'insieme delle disposizioni con ripetizione di classe 3 delle unità della popolazione; ogni campione $\mathbf{s} \in \mathcal{S}$ è quindi rappresentabile come terna ordinata (i_1, i_2, i_3), dove i_1, i_2, i_3 sono rispettivamente la prima, la seconda, la terza unità del campione $(i_1, i_2, i_3 \in I_N)$.
– Il campione $\mathbf{s} = (i_1, i_2, i_3)$ ha probabilità $p(\mathbf{s}) = p_{i_1}p_{i_2}p_{i_3}$. In pratica vengono effettuate tre "prove indipendenti", in ognuna delle quali si seleziona un'unità campionaria in modo tale che l'unità i ha probabilità p_i di essere selezionata.
a. Verificare che lo stimatore (di Hansen-Hurwitz)

$$t_{HH} = \frac{1}{3N}\left\{\frac{y_{i_1}}{p_{i_1}} + \frac{y_{i_2}}{p_{i_2}} + \frac{y_{i_3}}{p_{i_3}}\right\}$$

è uno stimatore corretto della media della popolazione.
b. Se $r(\mathbf{s}) = \{i, j\}$, mostrare che

$$E[t_{HH}\,|\boldsymbol{y}(r(\mathbf{s}))] = \frac{1}{3N}\left\{\frac{y_i}{p_i} + \frac{y_j}{p_j} + \frac{y_i + y_j}{p_i + p_j}\right\}.$$

Suggerimento. $r(\mathbf{s}) = \{i, j\}$ significa che il campione è l'uno o l'altro tra (i, i, j), (i, j, i), (j, i, i) (ciascuno ha probabilità $p_i^2 p_j$), (i, j, j), (j, i, j), (j, j, i) (ciascuno ha probabilità $p_i p_j^2$).

13.7. Verificare la relazione (13.27).

Suggerimento. Se T è sufficiente completa, vi può essere solo uno stimatore corretto di θ che è funzione di T. Inoltre, una statistica sufficiente completa è anche minimale.

14

Stimatori lineari della media della popolazione

14.1 Stimatori lineari: aspetti introduttivi

In questo capitolo ci si occuperà principalmente dei problemi di stima della media della popolazione, $\mu_y = \sum_{i=1}^{N} y_i/N$. Si accennerà inoltre, molto brevemente, anche al problema della stima del totale, $\theta = \sum_{i=1}^{N} y_i = N\mu_y$. A meno che non venga espressamente detto il contrario, nel seguito si assumerà sempre che il disegno campionario sia non ordinato e senza ripetizioni. In caso contrario è sufficiente passare alla riduzione del disegno stesso.

La classe fondamentale di stimatori a cui si farà riferimento è quella degli stimatori lineari. L'idea di fondo su cui si basano gli stimatori lineari è semplicissima: poiché la media della popolazione è una combinazione lineare delle y_i per tutte le unità della popolazione, si adotta come sua stima una combinazione lineare delle y_i per le sole unità campionarie. Uno stimatore (di μ_y) è *lineare* se può essere scritto nella forma

$$t = c_{0\mathbf{s}} + \frac{1}{N}\sum_{i\in\mathbf{s}} c_{i\mathbf{s}} y_i. \tag{14.1}$$

Il termine $c_{0\mathbf{s}}$ è l'*intercetta* dello stimatore, mentre i numeri $c_{i\mathbf{s}}$, $i \in \mathbf{s}$, rappresentano i *pesi* delle unità campionarie nello stimatore (14.1). In termini intuitivi, il peso dell'unità i può essere pensato come il numero di unità della popolazione "rappresentate" dall'unità i del campione. I pesi $c_{i\mathbf{s}}$ delle unità campionarie $i \in \mathbf{s}$ possono dipendere in generale sia dalle unità campionarie che dal campione $\mathbf{s}$, nel senso che una stessa unità i può ricevere un peso diverso a seconda del campione $\mathbf{s}$ a cui appartiene. L'intercetta $c_{0\mathbf{s}}$ può in generale dipendere dal campione $\mathbf{s}$ selezionato.

Uno stimatore (di μ_y) è *lineare omogeneo* se è lineare con intercetta nulla, ossia se può essere scritto come

$$t = \frac{1}{N}\sum_{i\in\mathbf{s}} c_{i\mathbf{s}} y_i. \tag{14.2}$$

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_14,

Come in precedenza accennato, se si confrontano gli stimatori (14.1), (14.2) con il parametro della popolazione da stimare

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{14.3}$$

si vede che, prescindendo dalla eventuale presenza dell'intercetta, la (14.1), la (14.2) e la (14.3) hanno struttura del tutto simile. La differenza sostanziale è che in (14.3) ogni unità *della popolazione* compare con un peso pari a 1, mentre in (14.1) e (14.2) ogni unità *del campione* compare con un peso $c_{i\mathbf{s}}$, $i \in \mathbf{s}$. Questa osservazione permette di fornire un'interpretazione euristica dei pesi che compaiono negli stimatori (14.1) e (14.2): $c_{i\mathbf{s}}$ si può pensare come il *numero di unità della popolazione rappresentate dall'unità i del campione* $\mathbf{s}$.

Gli stimatori (14.1) e (14.2) possono anche essere scritti introducendo gli indicatori di presenza/assenza delle unità nel campione. Posto

$$\delta(i;\ \mathbf{s}) = \begin{cases} 1 \text{ se } i \in \mathbf{s} \\ 0 \text{ se } i \notin \mathbf{s} \end{cases}$$

uno stimatore lineare si può scrivere come

$$t = c_{0\mathbf{s}} + \frac{1}{N}\sum_{i=1}^{N} c_{i\mathbf{s}}\delta(i;\ \mathbf{s})y_i \tag{14.4}$$

mentre uno stimatore lineare omogeneo si può scrivere come

$$t = \frac{1}{N}\sum_{i=1}^{N} c_{i\mathbf{s}}\delta(i;\ \mathbf{s})y_i. \tag{14.5}$$

Esempio 14.1 (Media campionaria). Sia $I_N = \{1, \ldots, N\}$ una popolazione finita di numerosità N, e si consideri un disegno $(\mathcal{S},\ p(\cdot))$, non ordinato e senza ripetizioni. Detto $n(\mathbf{s})$ il numero di unità nel campione $\mathbf{s}$, la *media campionaria* (introdotta nel caso di disegno ssr) assume la forma

$$\begin{aligned} \overline{y}_{\mathbf{s}} &= \frac{1}{n(\mathbf{s})}\sum_{i\in\mathbf{s}} y_i \\ &= \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{N}{n(\mathbf{s})} y_i \end{aligned} \tag{14.6}$$

da cui si desume che si tratta di uno stimatore lineare omogeneo di μ_y. I pesi delle unità sono tutti uguali, e pari a $N/n(\mathbf{s})$. Questo tipo di peso è supportato da una semplice intuizione: ogni unità del campione "rappresenta" $N/n(\mathbf{s})$ unità della popolazione. Chiaramente, in questo caso i pesi dipendono dal campione $\mathbf{s}$, ma non dalle unità i.

Se il disegno campionario è ad ampiezza costante ($n(\mathbf{s}) = n$, come accade ad es. nel disegno ssr) la (14.6) si riduce a

$$\frac{1}{n}\sum_{i\in\mathbf{s}} y_i.$$

In questo caso speciale i pesi campionari sono tutti pari a N/n, e non dipendono né dalle unità i, né dal campione $\mathbf{s}$. □

Esempio 14.2 (Stimatore per quoziente). Data una popolazione finita $I_N = \{1, \ldots, N\}$ di numerosità N, assumiamo che siano note le modalità x_i assunte da una variabile $\mathcal{X}$ su tutte le unità della popolazione; di conseguenza, è anche nota la media $\mu_x = \sum_{i=1}^{N} x_i/N$ di $\mathcal{X}$ nella popolazione. Si consideri un disegno $(\mathcal{S},\ p(\cdot))$, non ordinato e senza ripetizioni, e si indichi al solito con $n(\mathbf{s})$ il numero di unità nel campione $\mathbf{s}$. Lo *stimatore per quoziente* di μ_y assume la forma

$$\widehat{\mu}_q = \frac{\overline{y}_\mathbf{s}}{\overline{x}_\mathbf{s}}\mu_x \tag{14.7}$$

dove $\overline{y}_\mathbf{s}$ è la media campionaria della y_i e $\overline{x}_\mathbf{s}$ è la media campionaria delle x_i. Se si riscrive lo stimatore per quoziente come

$$\widehat{\mu}_q = \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{N\mu_x}{n(\mathbf{s})\overline{x}_\mathbf{s}}\, y_i \quad \textit{per ciascun } i \in \mathbf{s} \tag{14.8}$$

si vede subito che si tratta di uno stimatore lineare omogeneo di μ_y. I pesi delle unità campionarie sono tutti uguali, e pari a

$$c_{i\mathbf{s}} = \frac{N\mu_x}{n(\mathbf{s})\overline{x}_\mathbf{s}}. \tag{14.9}$$

Essendo $N\mu_x$ l'ammontare del carattere ausiliario $\mathcal{X}$ nella popolazione e $n(\mathbf{s})\overline{x}_\mathbf{s} = \sum_{i\in\mathbf{s}} x_i$ l'ammontare di $\mathcal{X}$ nel campione, si evince subito la logica alla base dei pesi (14.9). Ogni unità del campione riceve un peso tanto più grande quanto più grande è l'ammontare di $\mathcal{X}$ nella popolazione rispetto all'ammontare di $\mathcal{X}$ nel campione.

Come messo in evidenza nel Capitolo 6, lo stimatore per quoziente è distorto se il disegno utilizzato è di tipo semplice senza ripetizione. Ciò non esclude che usato con altri disegni possa essere corretto. In particolare, se si usa il disegno di Midzuno-Lahiri (cfr. Cap. 12) con $p_i = x_i/(N\mu_x)$, si vede facilmente che $\widehat{\mu}_q$ è uno stimatore corretto di μ_y (cfr. Esercizio 14.2). □

Esempio 14.3 (Stimatore per regressione). Consideriamo ancora la situazione dell'Esempio 14.2. Con l'usuale simbologia, lo *stimatore per regressione* di μ_y assume la forma

$$\widehat{\mu}_{reg} = \overline{y}_\mathbf{s} - \widehat{b}_{y/x}(\overline{x}_\mathbf{s} - \mu_x) \tag{14.10}$$

dove $\widehat{b}_{y/x}$ è il coefficiente di regressione campionario di $\mathcal{Y}$ rispetto a $\mathcal{X}$:

$$\widehat{b}_{y/x} = \frac{\widehat{s}_{xy}}{\widehat{s}_x^2} = \frac{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})\,(y_i - \overline{y}_\mathbf{s})}{\frac{1}{n-1}\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})^2}.$$

È immediato verificare (Esercizio 14.1) che lo stimatore per regressione può essere riscritto nella forma

$$\widehat{\mu}_{reg} = \frac{1}{N}\sum_{i\in\mathbf{s}} N\left\{\frac{1}{n(\mathbf{s})} - \frac{\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})(\overline{x}_\mathbf{s} - \mu_x)}{\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})^2}\right\} y_i \qquad (14.11)$$

da cui si deduce che si tratta di uno stimatore lineare omogeneo di μ_y con pesi

$$c_{i\mathbf{s}} = N\left\{\frac{1}{n(\mathbf{s})} - \frac{\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})(\overline{x}_\mathbf{s} - \mu_x)}{\sum_{i\in\mathbf{s}}(x_i - \overline{x}_\mathbf{s})^2}\right\}. \qquad \square$$

Il valore atteso dello stimatore (14.1) è studiato nella seguente proposizione.

Proposizione 14.1. *Sia t è uno stimatore lineare. Si ha allora*

$$E[t] = E[c_{0\mathbf{s}}] + \frac{1}{N}\sum_{i=1}^{N} E[c_{i\mathbf{s}}\,\delta(i;\,\mathbf{s})]\,y_i. \qquad (14.12)$$

Inoltre, t è corretto se e solo se

$$E[c_{0\mathbf{s}}] = 0; \quad E[c_{i\mathbf{s}}\,\delta(i;\,\mathbf{s})] = 1 \;\; \textit{per ciascun } i = 1, \ldots, N. \qquad (14.13)$$

Dimostrazione. Per la dimostrazione della (14.12) basta usare l'espressione (14.4), dalla quale si ottiene subito la (14.12).

Per quanto riguarda la (14.13), osserviamo in primo luogo che t è corretto se e solo se

$$E[c_{0\mathbf{s}}] + \frac{1}{N}\sum_{i=1}^{N} E[c_{i\mathbf{s}}\,\delta(i;\,\mathbf{s})]y_i = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \textit{per ciascun } \boldsymbol{Y}_N \in \Omega_N$$

ovvero se e solo se

$$E[c_{0\mathbf{s}}] + \frac{1}{N}\sum_{i=1}^{N}(E[c_{i\mathbf{s}}\,\delta(i;\,\mathbf{s})] - 1)y_i = 0 \quad \textit{per ciascun } \boldsymbol{Y}_N \in \Omega_N. \qquad (14.14)$$

Chiaramente, la (14.14) può aver luogo se e solo se

$$E[c_{0\mathbf{s}}] = 0; \quad E[c_{i\mathbf{s}}\,\delta(i;\,\mathbf{s})] - 1 = 0 \;\; \textit{per ciascun } i = 1, \ldots, N$$

ossia se e solo se vale la (14.13). $\square$

In chiusura, qualche osservazione sul problema della stima dell'ammontare $\theta = \sum y_i = N\mu_y$ del carattere $\mathcal{Y}$ nella popolazione. Se t è uno stimatore lineare di μ_y, allora $\widehat{\theta} = Nt$ è uno stimatore lineare di θ. Le proprietà di $\widehat{\theta}$ possono facilmente ottenersi a partire da quelle di t. In particolare, si vede subito che $E[\widehat{\theta}] = NE[t]$, e che $\widehat{\theta} = Nt$ è uno stimatore corretto dell'ammontare se e solo se t è uno stimatore corretto della media μ_y.

14.2 Un sempreverde del campionamento: lo stimatore di Horvitz-Thompson

14.2.1 Definizione e proprietà di base

Lo stimatore di Horvitz-Thompson (cfr. Horvitz e Thompson (1952)) è probabilmente il più importante tra gli stimatori lineari della media della popolazione. Esso è definito come:

$$t_{HT} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} y_i \tag{14.15}$$

dove π_i è la probabilità di inclusione (del primo ordine) dell'unità i.

Dalla (14.15) è facile desumere alcune proprietà elementari di t_{HT}. In primo luogo, si tratta di uno stimatore lineare omogeneo di μ_y. I pesi sono pari a

$$c_{i\mathbf{s}} = \frac{1}{\pi_i} \quad \textit{per ciascun } \mathbf{s} \in \mathcal{S} \textit{ e } i = 1, \ldots, N \tag{14.16}$$

e quindi dipendono *solo* dalle unità, ma *non* dai campioni. Una stessa unità, in campioni diversi, riceve sempre lo stesso peso. È anche da rimarcare che il peso dell'unità i è il reciproco della sua probabilità di inclusione, ovvero il reciproco della probabilità che tra le unità del campione vi sia i. Il peso $1/\pi_i$ della generica unità i è spesso denominato *coefficiente di riporto all'universo*, ed è inversamente proporzionale alla probabilità di selezione dell'unità i. Ciò significa che se per esempio $\pi_i = 0.02$ allora $1/\pi_i = 50$, e nello stimatore (14.16) è come se l'unità i rappresentasse 50 unità della popolazione. Tale ponderazione delle osservazioni campionarie propria dello stimatore di Horvitz-Thompson rappresenta un correttivo della diversa probabilità che le varie unità hanno di figurare nel campione. In linea di principio, tendono a essere più rappresentate nel campione le modalità y_i delle unità i che possiedono una probabilità di inclusione più alta, e ad essere meno rappresentate nel campione le modalità y_i delle unità i che possiedono una probabilità di inclusione più bassa. I pesi dello stimatore di Horvitz-Thompson rappresentano una forma di "bilanciamento" di questo fatto.

Le proprietà più semplici dello stimatore di Horvitz-Thompson sono riportate nella successiva Proposizione 14.2. Prima di enunciarla (e dimostrarla) osserviamo che, esattamente come per gli stimatori lineari del paragrafo

precedente, lo stimatore t_{HT} si può scrivere nella forma

$$t_{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} \delta(i;\ \mathbf{s})\, y_i \tag{14.17}$$

dove $\delta(i;\ \mathbf{s})$ è il solito indicatore di presenza/assenza dell'unità i nel campione.

Proposizione 14.2. *Se il disegno campionario* $(\mathcal{S}, p(\cdot))$ *è tale che* $\pi_i > 0$ *per tutte le unità della popolazione, lo stimatore* t_{HT} (14.15) *possiede le seguenti proprietà:*

- *è uno stimatore corretto della media della popolazione:*

$$E[t_{HT}] = \mu_y \quad \textit{qualunque sia } \boldsymbol{Y}_N \in \Omega_N\,; \tag{14.18}$$

- *è l'unico stimatore lineare omogeneo corretto di* μ_y *i cui pesi dipendono solo dall'unità* i *ma non dal campione* $\mathbf{s}$ *(ovvero sono tali che* $c_{i\mathbf{s}} = c_i$*);*
- *la sua varianza assume la forma*

$$V(t_{HT}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij} \tag{14.19}$$

con

$$\Delta_{ij} = \pi_{ij} - \pi_i \pi_j;\ \ i,\, j = 1,\, \ldots,\, N. \tag{14.20}$$

Dimostrazione. Per la (14.18) basta osservare che

$$\begin{aligned} E[t_{HT}] &= E\left[\frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} \delta(i;\ \mathbf{s})\, y_i\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} E[\delta(i;\ \mathbf{s})]\, y_i \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} \pi_i y_i \\ &= \mu_y. \end{aligned} \tag{14.21}$$

Per quanto riguarda la proprietà di unicità dello stimatore di Horvitz-Thompson, uno stimatore lineare omogeneo della forma

$$t = \frac{1}{N} \sum_{i=1}^{N} c_i\, \delta(i;\ \mathbf{s})\, y_i \tag{14.22}$$

è corretto se e solo se

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} y_i &= E\,[t] \\ &= \frac{1}{N} E\left[\sum_{i=1}^{N} c_i\,\delta(i;\ \mathbf{s})\,y_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} c_i\,E[\delta(i;\ \mathbf{s})]\,y_i \\ &= \frac{1}{N}\sum_{i=1}^{N} c_i\,\pi_i\,y_i. \end{aligned} \tag{14.23}$$

Ma la (14.23) può aver luogo se e solo se

$$c_i\,\pi_i = 1 \quad \textit{per ciascun } i = 1,\ \ldots,\ N \tag{14.24}$$

ossia se e solo se $c_1 = 1/\pi_i$. In questo caso lo stimatore (14.22) si riduce allo stimatore di Horvitz-Thompson di μ_y.

Infine, per quanto riguarda la varianza dello stimatore di Horvitz-Thompson, si ha

$$\begin{aligned} V(t_{HT}) &= V\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}\,\delta(i;\ \mathbf{s})\,y_i\right) \\ &= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2 V(\delta(i;\ \mathbf{s})) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,C(\delta(i;\ \mathbf{s}),\ \delta(j;\ \mathbf{s}))\right\} \\ &= \frac{1}{N^2}\left\{\sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2 (\pi_i - \pi_i^2) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,(\pi_{ij} - \pi_i\pi_j)\right\} \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,(\pi_{ij} - \pi_i\pi_j). \end{aligned}$$

□

Se il disegno campionario è ad ampiezza effettiva costante, si può anche fornire un'espressione alternativa della varianza dello stimatore di Horvitz-Thompson della media della popolazione, utile soprattutto per stimare la varianza stessa. Formalmente vale la seguente proposizione (cfr. Yates e Grundy (1953)).

Proposizione 14.3. *Se il disegno campionario è ad ampiezza effettiva costante n, la varianza dello stimatore t_{HT} si può scrivere come:*

$$V(t_{HT}) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2(\pi_i\pi_j-\pi_{ij})$$
$$= -\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\Delta_{ij} \qquad (14.25)$$

con Δ_{ij} dato dalla (14.20).

Dimostrazione. È sufficiente osservare che

$$\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2(\pi_i\pi_j-\pi_{ij})$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}\right)^2$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} - \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{y_i}{\pi_i}\right)^2\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)$$
$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

in quanto

$$\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j) = \sum_{j=1}^{N}\pi_{ij} - \pi_i\sum_{j=1}^{N}\pi_j$$
$$= n\pi_i - n\pi_i$$
$$= 0.$$

□

14.2.2 Costruzione dello stimatore di Horvitz-Thompson per disegni campionari "semplici"

Lo stimatore di Horvitz-Thompson comprende come casi particolari molti degli stimatori elementari studiati nei capitoli precedenti.

Esempio 14.4 (Media campionaria nel disegno ssr). Data una popolazione $I_N = \{1, \ldots, N\}$ di numerosità N, supponiamo di selezionare da essa un campione $\mathbf{s}$ di numerosità n mediante disegno ssr. Come già visto nel Capitolo 12, le probabilità di inclusione del primo ordine sono $\pi_i = n/N$ per

tutte le unità della popolazione. Lo stimatore di Horvitz-Thompson assume pertanto la forma:

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{n/N} y_i \\ &= \frac{1}{n} \sum_{i \in \mathbf{s}} y_i \\ &= \overline{y}_{\mathbf{s}} \end{aligned}$$

ossia si riduce all'usuale media campionaria. □

Esempio 14.5 (Disegno campionario stratificato). Supponiamo che la popolazione sia suddivisa in M strati, rispettivamente di $N_1, \ldots, N_M$ unità (elementari). Indichiamo poi, al solito, con $\mathbf{s}_g$ $(g = 1, \ldots, M)$ il sottocampione di n_g unità (elementari) selezionate dallo strato g-mo $(g = 1, \ldots, M)$ mediante disegno ssr. Le probabilità di inclusione del primo ordine, calcolate nel Capitolo 12, sono eguali a $\pi_{(g)i} = n_g/N_g$ per ciascuna unità dello strato g-mo, e per tutti gli strati g $(g = 1, \ldots, M)$.

Lo stimatore di Horvitz-Thompson assume, con la notazione usata nel caso stratificato, la forma

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{g=1}^{M} \sum_{i \in \mathbf{s}_g} \frac{1}{\pi_{(g)i}} y_{gi} \\ &= \sum_{g=1}^{M} \frac{N_g}{N} \left\{ \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi} \right\} \\ &= \sum_{g=1}^{M} w_g \overline{y}_g \\ &= \widehat{\mu}_{str} \end{aligned}$$

ovvero si riduce al "solito" stimatore introdotto nel Capitolo 7. □

Esempio 14.6 (Disegno campionario a grappolo "semplice"). Supponiamo che la popolazione sia composta da M grappoli, rispettivamente di $N_1, \ldots, N_M$ unità elementari. Nel caso di disegno a grappolo (con uguali probabilità di selezione dei grappoli), la probabilità di inclusione del primo ordine dell'unità elementare i del grappolo g, calcolata nel Capitolo 12, è pari a $\pi_{(g)i} = m/M$, dove m è il numero di grappoli selezionato (mediante disegno ssr).

Detto, come al solito, $\mathbf{g}_m$ il campione di grappoli selezionato, i dati statistici osservati saranno $\{y_{gi};\quad i = 1, \ldots, N_g;\ g \in \mathbf{g}_m\}$. Lo stimatore di

Horvitz-Thompson assume pertanto la forma

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \sum_{i=1}^{N_g} \frac{1}{\pi_{(g)i}} y_{gi} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M \frac{N_g}{N} \left\{ \frac{1}{N_g} \sum_{i=1}^{N_g} y_{gi} \right\} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \mu_{yg} \\ &= \widehat{\mu}_{gr} \end{aligned}$$

e quindi è null'altro che lo stimatore introdotto nel Capitolo 9. □

Esempio 14.7 (Disegno campionario a due stadi semplici). Supponiamo ancora che la popolazione sia composta da M grappoli, rispettivamente di $N_1, \ldots, N_M$ unità elementari, e che il disegno di selezione delle unità sia a due stadi semplici. La probabilità di inclusione del primo ordine dell'unità elementare i del grappolo g, calcolata nel Capitolo 12, è pari a $\pi_{(g)i} = m/M\, n_g/N_g$, dove m è il numero di grappoli selezionato (mediante disegno ssr) al primo stadio, e n_g è il numero di unità elementari selezionate al secondo stadio dal grappolo g-mo, a sua volta selezionato al primo stadio.

Detti, come al solito, $\mathbf{g}_m$ il campione di grappoli selezionato al primo stadio, e $\mathbf{s}_g, g \in \mathbf{g}_m$ i campioni di unità elementari selezionate al secondo stadio di campionamento, i dati statistici osservati saranno $\{y_{gi};\ \ i \in \mathbf{s}_g;\ g \in \mathbf{g}_m\}$. Lo stimatore di Horvitz-Thompson assume quindi la forma:

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \sum_{i \in \mathbf{s}_g} \frac{1}{\pi_{(g)i}} y_{gi} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} M \frac{N_g}{N} \left\{ \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi} \right\} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \overline{y}_g \\ &= \widehat{\mu}_{2st} \end{aligned}$$

e quindi è lo stimatore introdotto nel Capitolo 11. □

14.2.3 Stima della varianza dello stimatore di Horvitz-Thompson: risultati esatti

Il problema della stima della varianza dello stimatore di Horvitz-Thompson è un problema di notevole importanza, anche se purtroppo, per molti aspetti, non possiede una soluzione del tutto soddisfacente. Una prima idea per costruire uno stimatore non distorto della (14.19) consiste nell'usare idee simili a quelle già utilizzate per la stima della media della popolazione. Formalmente, la media della popolazione $\mu_y = \sum_i y_i/N$ è null'altro che una funzione

lineare di $y_1, \ldots, y_N$; per stimarla si è fatto ricorso ad una funzione *lineare* dei dati campionari. La varianza (14.19) è una funzione quadratica di $y_1, \ldots, y_N$; per stimarla è allora naturale utilizzare una funzione *quadratica* dei dati campionari. Nella successiva Proposizione 14.4 viene costruito lo stimatore di $V(t_{HT})$ originariamente proposto in Horvitz e Thompson (1952). La condizione essenziale su cui si basa la Proposizione 14.4 è che tutte le coppie i, j di unità distinte abbiano *probabilità di inclusione del secondo ordine positiva*. Se $\pi_{ij} = 0$, nessun campione conterrà simultaneamente le unità i e j, e quindi non sarà possibile costruire un stimatore corretto di $V(t_{HT})$. È quello che accade, per esempio, nel disegno sistematico, per il quale si rinvia all'Esercizio 14.3.

Proposizione 14.4. *Sotto la condizione $\pi_{ij} > 0$ per ogni coppia di unità i,j (condizione di misurabilità del disegno) uno stimatore non distorto della varianza di uno stimatore lineare è il seguente*

$$\widehat{V}_{HT}(t_{HT}) = \frac{1}{N^2} \sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}} \,. \tag{14.26}$$

Dimostrazione. Basta osservare che

$$\begin{aligned} E[\widehat{V}_{HT}(t_{HT})] &= \frac{1}{N^2} E \left[\sum_{i \in \mathbf{s}} \sum_{j \in \mathbf{s}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}} \right] \\ &= \frac{1}{N^2} E \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}} \delta(i;\, \mathbf{s}) \delta(j;\, \mathbf{s}) \right] \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}} E[\delta(i;\, \mathbf{s}) \delta(j;\, \mathbf{s})] \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}} \pi_{ij} \\ &= V(t_{HT}) \,. \end{aligned} \tag{14.27}$$

□

Lo stimatore (14.26), pur essendo corretto, ha una caratteristica poco invidiabile: può assumere valori negativi. Per alcuni particolari valori y_i, e per qualche particolare campione $\mathbf{s}$, potrebbe aversi $\widehat{V}_{HT}(t_{HT}) < 0$.

Esempio 14.8. Si consideri una popolazione finita di $N = 3$ unità, dalla quale si seleziona un campione di $n = 2$ unità. Il disegno campionario è qui sotto specificato:

$$\mathcal{S} = \{\{1, 2\},\ \{1, 3\},\ \{2, 3\}\}$$
$$p(\{1, 2\}) = 0.1,\ p(\{1, 3\}) = 0.45,\ p(\{2, 3\}) = 0.45.$$

Le probabilità di inclusione del primo e del secondo ordine sono rispettivamente eguali a:

$$\pi_1 = 0.55,\ \pi_2 = 0.55,\ \pi_3 = 0.9;\ \ \pi_{12} = 0.1,\ \pi_{13} = 0.45,\ \pi_{23} = 0.45.$$

Se il campione selezionato è $\{1, 2\}$, lo stimatore (14.26) è eguale a

$$\begin{aligned}\widehat{V}_{HT}(t_{HT}) &= \frac{1}{9}\left\{\frac{y_1^2}{\pi_1^2}\frac{\pi_1-\pi_1^2}{\pi_1} + \frac{y_2^2}{\pi_2^2}\frac{\pi_2-\pi_2^2}{\pi_2} + 2\frac{y_1}{\pi_1}\frac{y_2}{\pi_2}\frac{\pi_{12}-\pi_1\pi_2}{\pi_{12}}\right\} \\ &= \frac{1}{9}\left\{y_1^2\left(\frac{1}{\pi_1^2}-\frac{1}{\pi_1}\right) + y_2^2\left(\frac{1}{\pi_2^2}-\frac{1}{\pi_2}\right) + 2y_1y_2\left(\frac{1}{\pi_1\pi_2}-\frac{1}{\pi_{12}}\right)\right\} \\ &= \frac{1}{9}\left\{1.49\,y_1^2 + 1.49\,y_1^2 - 13.38\,y_1y_2\right\}. \qquad (14.28)\end{aligned}$$

In particolare, per $y_1 = y_1 = 9$ la (14.28) è pari a –95.22. □

A peggiorare le cose, c'è anche il fatto che in generale non è agevole fornire una condizione sufficiente sul disegno campionario che assicuri la non negatività dello stimatore (14.26).

Per le ragioni sopra menzionate, è opportuno cercare di costruire uno stimatore alternativo a (14.26). Un approccio promettente consiste nell'utilizzare l'espressione (14.25), che ovviamente vale solo per disegni ad ampiezza effettiva costante. Ciò porta allo stimatore di Yates-Grundy (14.29) (cfr. Yates e Grundy (1953)).

Proposizione 14.5. *Se il disegno campionario è a dimensione effettiva costante, e se $\pi_{ij} > 0$ per ogni coppia i, j di unità distinte uno stimatore corretto di $V(t_{HT})$ (14.25) è dato da*

$$\widehat{V}_{YG}(t_{HT}) = \frac{1}{2N^2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}. \qquad (14.29)$$

Dimostrazione. La correttezza dello stimatore (14.29) si dimostra immediatamente osservando che:

$$\begin{aligned}E[\widehat{V}_{YG}(t_{HT})] &= E\left[\frac{1}{2N^2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right] \\ &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}E[\delta(i;\mathbf{s})\delta(j;\mathbf{s})] \\ &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}.\end{aligned}$$

□

L'espressione (14.29) dello stimatore di Yates-Grundy di $V(t_{HT})$ suggerisce una semplice condizione sufficiente che ne garantisce la non negatività. *Se:*

$$\pi_{ij} \leqslant \pi_i \pi_j \quad \begin{array}{l} \textit{per tutte le coppie di unità distinte} \\ i,\ j \textit{ della popolazione} \end{array} \tag{14.30}$$

allora lo stimatore di Yates-Grundy (14.29) *è sempre non negativo.*

Esempio 14.9. Si consideri ancora l'Esempio 14.8, in cui $\widehat{V}_{HT}(t_{HT})$ può essere negativo. Essendo

$$\pi_1 \pi_2 = 0.3025,\ \pi_1 \pi_3 = \pi_2 \pi_3 = 0.495$$

la condizione (14.30) è soddisfatta, e lo stimatore $\widehat{V}_{YG}(t_{HT})$ assume solo valori non negativi. □

14.2.4 Stima della varianza dello stimatore di Horvitz-Thompson: risultati approssimati

I metodi di approssimazione delle probabilità di inclusione del secondo ordine permettono di fornire utili approssimazioni della varianza dello stimatore di Horvitz-Thompson, così come utili espressioni per suoi stimatori approssimati.

Se per le probabilità di inclusione del secondo ordine si usa la più semplice approssimazione sviluppata nella Sezione 12.7

$$\pi_{ij} \approx \pi_i\, \pi_j \left(1 - \frac{(1-\pi_i)\,(1-\pi_j)}{d}\right)$$

si ha, sempre in via approssimata,

$$\Delta_{ij} \approx -\frac{\pi_i\,(1-\pi_i)\,\pi_j\,(1-\pi_j)}{d}$$

con $d = \sum_i \pi_i\,(1-\pi_i)$. Pertanto, la (14.25) si scrive, in via approssimata, come

$$V(t_{HT}) \approx \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i\,(1-\pi_i)\,\pi_j\,(1-\pi_j)}{d}. \tag{14.31}$$

Posto

$$A = \sum_{i=1}^{N} y_i\, \frac{1-\pi_i}{d}$$

è facile verificare (Esercizio 14.16) che

$$\begin{aligned} &\sum_{i=1}^{N} \left(\frac{y_i}{\pi_i} - A\right)^2 \frac{\pi_i\,(1-\pi_i)}{d} \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i\,(1-\pi_i)}{d}\, \frac{\pi_j\,(1-\pi_j)}{d} \end{aligned} \tag{14.32}$$

da cui discende, in forza della (14.31), che

$$V(t_{HT}) \approx \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{y_i}{\pi_i} - A \right)^2 \pi_i \, (1 - \pi_i). \tag{14.33}$$

La costruzione di uno stimatore approssimativamente corretto della (14.33) procede senza particolari difficoltà. Tenendo infatti conto che stimatori corretti di $\sum_i y_i \, (1 - \pi_i)$ e di $\sum_i \pi_1 (1 - \pi_i)$ sono rispettivamente

$$\sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i} \, (1 - \pi_i), \quad \sum_{i \in \mathbf{s}} (1 - \pi_i)$$

come stimatore (distorto ma approssimativamente corretto) di A si può far riferimento al seguente

$$\widehat{A} = \sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i} \, (1 - \pi_i) \Bigg/ \sum_{i \in \mathbf{s}} (1 - \pi_i). \tag{14.34}$$

Come stimatore (distorto ma approssimativamente corretto) dell'espressione approssimata (14.33) si può quindi usare il seguente

$$\widehat{V}_{AP1}(t_{HT}) = \frac{1}{N^2} \sum_{i \in \mathbf{s}} \left(\frac{y_i}{\pi_i} - \widehat{A} \right)^2 (1 - \pi_i) \tag{14.35}$$

con $\widehat{A}$ dato dalla (14.34). Lo stimatore (14.35) è sostanzialmente equivalente ad uno proposto da Deville (vds. Tillé (2006), p. 141).

Considerazioni simili si possono fare se si usa l'approssimazione $\pi_{ij} \approx \pi_{ij}^a = \pi_i \, \pi_j - c_i \, c_j$ (in genere migliore della precedente), sempre sviluppata nella Sezione 12.7. Essendo $\Delta_{ij} \approx \delta_{ij}^a = -c_i \, c_j$, e posto $C = c_1 + \cdots + c_N$, si ha anzitutto

$$\begin{aligned} V(t_{HT}) &\approx \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 c_i \, c_j \\ &= \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{C \, y_i}{\pi_i} - \frac{C \, y_j}{\pi_j} \right)^2 c_i \, c_j \end{aligned} \tag{14.36}$$

da cui, posto

$$B = \sum_{i=1}^{N} \frac{y_i}{\pi_i} \, c_i$$

si ottiene

$$V(t_{HT}) \approx \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{C \, y_i}{\pi_i} - B \right)^2 \frac{c_i}{C}. \tag{14.37}$$

Come stimatore corretto di B si può usare

$$\widehat{B} = \sum_{i \in \mathbf{s}} \frac{y_i}{\pi_i^2} c_i$$

da cui, con lo stesso tipo di ragionamento già svolto in precedenza, si perviene allo stimatore (distorto ma approssimativamente corretto) della (14.37)

$$\widehat{V}_{AP2}(t_{HT}) = \frac{1}{N^2} \sum_{i \in \mathbf{s}} \left(\frac{C\, y_i}{\pi_i} - \widehat{B} \right)^2 \frac{c_i}{C\, \pi_i}. \tag{14.38}$$

Naturalmente, per la validità delle espressioni approssimate di $V(t_{HT})$ fornite nella presente sezione, e per i relativi stimatori, vale quanto già detto per le approssimazioni delle probabilità di inclusione del secondo ordine. Le approssimazioni (14.33), (14.37), e i corrispondenti stimatori (14.35), (14.38), forniscono risultati tanto migliori quanto più elevata è l'entropia del disegno campionario utilizzato.

Per un trattamento più esaustivo della costruzione di stimatori approssimati della varianza dello stimatore di Horvitz-Thompson si rinvia al volume di Tillé (2006), e all'articolo di Matei e Tillé (2005).

14.2.5 Stimatore di Horvitz-Thompson dell'ammontare di un carattere

Nella proposizione 14.2 si è dimostrato che lo stimatore di Horvitz-Thompson t_{HT} è uno stimatore corretto della media della popolazione, con varianza data dalla (14.19). È immediato estendere tale risultato all'ammontare totale di un carattere in una popolazione. Formalmente, poiché l'ammontare è dato dal prodotto tra la numerosità della popolazione e la media della popolazione stessa

$$\theta = \sum_{i=1}^{N} y_i = N \mu_y, \tag{14.39}$$

lo stimatore di Horvitz-Thompson dell'ammontare si ottiene moltiplicando N per lo stimatore di Horvitz-Thompson della media

$$\widehat{\theta}_{HT} = N\, t_{HT} = \sum_{i \in \mathbf{s}} \frac{1}{\pi_i}\, y_i. \tag{14.40}$$

Dai risultati ottenuti per lo stimatore di Horvitz-Thompson della media è facile trarre i seguenti risultati.

Proposizione 14.6. *Lo stimatore* $\widehat{\theta}_{HT}$ *dato dalla* (14.40) *è uno stimatore corretto del totale della popolazione*

$$E(\widehat{\theta}_{HT}) = t \tag{14.41}$$

con varianza pari a

$$V(\widehat{\theta}_{HT}) = \sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{14.42}$$

Uno stimatore corretto della (14.42) *è*

$$\widehat{V}_{HT}(\widehat{\theta}_{HT}) = \sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}\,\frac{\Delta_{ij}}{\pi_{ij}}. \tag{14.43}$$

Se poi il disegno campionario è ad ampiezza effettiva costante, la varianza di $\widehat{\theta}_{HT}$ *può alternativamente scriversi come*

$$V(\widehat{\theta}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2(\pi_i\pi_j - \pi_{ij}) \tag{14.44}$$

e un suo stimatore corretto è il seguente

$$\widehat{V}_{YG}(\widehat{\theta}_{HT}) = \frac{1}{2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}. \tag{14.45}$$

Dimostrazione. È una immediata conseguenza delle Proposizioni 14.2, 14.4, 14.5. □

14.2.6 Ruolo delle probabilità di inclusione sull'efficienza dello stimatore di Horvitz-Thompson nei disegni ad ampiezza effettiva costante

A determinare la varianza dello stimatore di Horvitz-Thompson, e quindi la sua efficienza, concorrono ovviamente tre elementi:

- il disegno campionario;
- la forma funzionale dell stimatore;
- il vettore $\boldsymbol{Y}_N$ delle modalità del carattere sulle unità della popolazione.

Mentre c è incognito allo statistico, e sostanzialmente "deciso dalla natura", gli altri due elementi sono oggetto di una precisa scelta da parte dello statistico. È quindi di interesse studiare quale tipo di disegno, se esiste, massimizza l'efficienza dello stimatore t_{HT}, ossia minimizza la sua varianza. Il ragionamento è semplice. Se il disegno campionario è ad ampiezza effettiva costante n, allora t_{HT} ha varianza nulla, e quindi efficienza massima, quando le probabilità di inclusione risultano proporzionali ai valori della variabile di interesse $\mathcal{Y}$: $\pi_i = cost\, y_i$, con *cost* costante opportuna. Per verificare ciò, osserviamo che se $\pi_i = cost\, y_i$ e se tutti i campioni hanno lo stesso numero n di

unità distinte, dalla relazione $\pi_1 + \cdots + \pi_N = n$ deve aversi:

$$\begin{aligned} n &= \sum_{i=1}^{N} \pi_i \\ &= cost \sum_{i=1}^{N} y_i \\ &= cost\, N\, \mu_y \end{aligned}$$

da cui si ricava $cost = n/(N\mu_y)$, e quindi

$$\pi_i = \frac{n}{N\mu_y} y_i \quad per\ ciascun\ i = 1, \ldots, N. \tag{14.46}$$

Con probabilità di inclusione del primo ordine (14.46), ed *ampiezza effettiva costante* n, lo stimatore di Horvitz-Thompson di μ_y diviene pari a

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{i=1}^{N} \frac{N\mu_y}{n y_i} y_i\, \delta(i;\, \mathbf{s}) \\ &= \frac{\mu_y}{n} \sum_{i=1}^{N} \delta(i;\, \mathbf{s}) \\ &= \frac{\mu_y}{n}\, n \\ &= \mu_y \quad per\ ciascun\ campione\ \mathbf{s}. \end{aligned} \tag{14.47}$$

La (14.47) mostra che lo stimatore t_{HT} è identicamente uguale alla media μ_y della popolazione, e quindi, in forza della sua correttezza, la sua varianza deve essere uguale a 0. Scegliere probabilità di inclusione del primo ordine (14.46), ed una numerosità campionaria effettiva costante, rende quindi lo stimatore di Horvitz-Thompson esattamente uguale al parametro da stimare. Si osservi che sul piano intuitivo le probabilità di inclusione "ottime" (14.46) sono "grandi" per le unità con valori y_i "grandi", mentre sono "piccole" per le unità con valori y_i "piccoli".

Dato che i valori di $\mathcal{Y}$ sono incogniti, le probabilità di inclusione "ottimali" (14.46) non possono essere calcolate. Tuttavia, se sono noti i valori x_i, $i = 1, \ldots, N$ di una variabile ausiliaria $\mathcal{X}$ correlata con la variabile di interesse $\mathcal{Y}$, si può pensare di sfruttare questi valori per una buona scelta delle probabilità di inclusione. Ad es., se $\mathcal{X}$ è positivamente correlata con $\mathcal{Y}$, si può pensare di assegnare probabilità di inclusione "grandi" alle unità con valori x_i "grandi" (perché questi sono presumibilmente associati a valori y_i "grandi"), e probabilità di inclusione "piccole" alle unità con valori x_i "piccoli" (perché questi sono presumibilmente associati a valori y_i "piccoli"). Un esempio (uno dei molti!) di disegno campionario che soddisfa questo requisito è il disegno di Midzuno-Lahiri con $p_i = x_i / \sum_{j=1}^{N} x_j$ (cfr. Capitolo 12).

Una scelta più precisa, e migliore, si può effettuare nel caso, importante nelle applicazioni, in cui tra le variabili di interesse $\mathcal{Y}$ e ausiliaria $\mathcal{X}$ sussista una relazione di approssimata proporzionalità:

$$\frac{y_i}{x_i} \approx \textit{costante}, \quad i = 1, \ldots, N. \tag{14.48}$$

In questo caso scegliere probabilità di inclusione del primo ordine proporzionali alle x_i significa, in forza della (14.48), che esse sono anche approssimativamente proporzionali alle y_i. Detta μ_x la media di $\mathcal{X}$ nella popolazione, se vale la (14.48) una buona scelta delle probabilità di inclusione del primo ordine consiste nel porle eguali a

$$\pi_i = \frac{nx_i}{\sum_{j=1}^{N} x_j} = \frac{nx_i}{N\,\mu_x}. \tag{14.49}$$

Il rapporto $x_i / \sum_{j=1}^{N} x_j$ rappresenta la dimensione relativa dell'unità i.

Un disegno di campionamento in cui le probabilità di inclusione del primo ordine sono scelte con il criterio (14.49) è detto πpps (*inclusion probabilities proportional to size*).

L'utilizzo di un disegno campionario a probabilità variabili, e che assegna ad ogni unità della popolazione una probabilità di inclusione proporzionale al valore di una variabile ausiliaria (nota per tutte le unità della popolazione) dovrebbe garantire, qualora si sia non lontani dalla (14.48), una buona efficienza dello stimatore di Horvitz-Thompson. Usare a livello di definizione del disegno campionario l'informazione derivante dalla conoscenza della variabile ausiliaria $\mathcal{X}$ indurrà una diminuzione della varianza degli stimatori dei parametri di interesse, conducendo alla definizione di stimatori più efficienti rispetto a quelli ottenibili da una selezione mediante disegno semplice.

Chiaramente per ogni unità della popolazione si deve avere $\pi_i \leqslant 1$. Notiamo che tale condizione è certamente soddisfatta per $n = 1$; se $n > 1$, e in corrispondenza di un valore elevato di x_i, la (14.49) potrebbe portare a valori $\pi_i > 1$ per qualche unità della popolazione. In tali condizioni l'unità dovrà essere inclusa con certezza nel campione. Formalmente, si porrà $\pi_i = 1$ per tutte le unità i tali che $nx_i \geqslant \sum_{j=1}^{N} x_j$, e si ricalcoleranno le probabilità di inclusione delle restanti unità come

$$\pi_i = (n - n_A)\frac{x_i}{\sum_{j=1_{j\notin A}}^{N} x_j}; \quad i = 1, \ldots, N;\ i \notin A$$

dove A è l'insieme delle unità della popolazione tali che $nx_i \geqslant \sum_{j=1}^{N} x_j$, e n_A il numero di tali unità.

L'efficienza dello stimatore di Horvitz-Thompson ha destato parecchio dibattito nella letteratura statistica. Un esempio pittoresco, e illuminante, sul perché in alcune circostanze esso possa condurre a risultati pessimi è il seguente, dovuto a Basu (1971).

Esempio 14.10 (Elefanti del circo; Basu (1971)). Il proprietario di un circo deve imbarcare i suoi 50 elefanti, e quindi ha bisogno di una stima del loro peso totale. Poiché pesare un elefante è un'operazione lunga e difficile, il proprietario decide di pesare solo un elefante. Quale dei 50 scegliere? Gli elefanti erano stati pesati tre anni addietro, e dal relativo elenco dei pesi il proprietario del circo scopre che Sambo, un elefante di taglia media, aveva un peso praticamente pari al peso medio dei 50 elefanti. Poiché nel frattempo gli elefanti non sono cambiati di molto, il proprietario ritiene che Sambo abbia ancora un peso grosso modo pari alla media dei 50 elefanti del circo. Pertanto, egli propone di prendere Sambo, pesarlo, e stimare il peso totale di tutti gli elefanti come segue:

$$\textit{peso Sambo} \times 50.$$

Lo statistico del circo, però, è molto critico verso questa scelta. In primo luogo, egli afferma che non è corretto usare un disegno campionario ragionato, che dà probabilità di inclusione 1 a Sambo, e 0 a tutti gli altri elefanti. Meglio invece un disegno campionario in cui Sambo ha probabilità 99/100 di essere scelto, e ciascuno degli altri 49 ha probabilità 1/4900. Una volta utilizzato, questo disegno campionario seleziona (naturalmente!) Sambo. Il proprietario del circo, contento del risultato, pensa di stimare il peso totale dei 50 elefanti moltiplicando per 50 il peso di Sambo. Ma anche qui lo statistico ha da ridire. A suo avviso è meglio usare lo stimatore di Horvitz-Thompson, che possiede molte belle proprietà. Se ad essere scelto è Sambo, il peso totale dei 50 elefanti è stimato pari a:

$$\textit{peso Sambo} \times \frac{100}{99}.$$

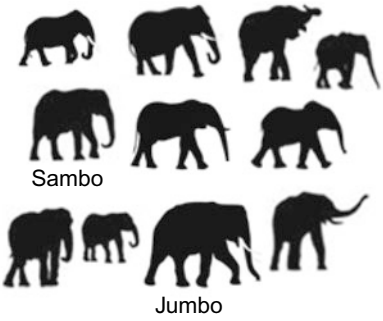
Il proprietario del circo è molto perplesso. Per questa ragione chiede allo statistico come avrebbe stimato il peso totale dei 50 elefanti se ad essere selezionato fosse stato Jumbo, decisamente una taglia forte. Anche qui la risposta dello statistico è netta. Usando lo stimatore di Horvitz-Thompson, il peso totale dei 50 elefanti è stimato pari a:

$$\textit{peso Jumbo} \times 4900.$$

Il diverso comportamento del proprietario del circo e dello statistico è illustrato in Fig. 14.1.

Inutile dire che subito dopo aver formulato la sua proposta lo statistico ha dovuto cercare un altro lavoro. □

L'Esempio 14.10 è stato interpretato come "distruttivo" per lo stimatore di Horvitz-Thompson. In realtà, esso mostra una cosa piuttosto nota nella pratica applicativa: lo stimatore di Horvitz-Thompson può dare risultati assurdi se le probabilità di inclusione del primo ordine non hanno nessun legame con i valori y_i. Nell'Esempio 14.10 l'uso "ottimale" dello stimatore di Horvitz-Thompson richiederebbe di dare agli elefanti probabilità di inclusione (e di selezione, visto che ci si basa su un campione di $n = 1$ unità)



Quanto pesano i 50 elefanti del circo?

Idea del proprietario del circo
Scegliere l'elefante Sambo

Stima peso totale = peso Sambo × 50
ERRORE DI STIMA "PICCOLO"

Idea dello statistico
Scegliere "casualmente" uno degli elefanti, in modo che Sambo abbia probabilità $\frac{99}{100}$ di essere scelto, e ciascuno degli altri probabilità $\frac{1}{4900}$

Se è selezionato Sambo

Stima peso totale = peso Sambo × $\frac{100}{99}$
GRAVE SOTTOSTIMA

Se è selezionato Jumbo

Stima peso totale = peso Jumbo × 4900
GRAVE SOVRASTIMA

Fig. 14.1 Esempio degli elefanti del circo

proporzionali al peso degli elefanti stessi. Jumbo, l'elefante taglia forte, dovrebbe avere la probabilità di inclusione più alta. Sambo, l'elefante taglia media, dovrebbe avere una probabilità di inclusione decisamente più piccola. Naturalmente, non essendo disponibili i pesi attuali degli elefanti, ma solo quelli di tre anni addietro, un buon disegno campionario da usare in coppia con lo stimatore di Horvitz-Thompson dovrebbe prevedere probabilità di inclusione degli elefanti proporzionali al loro peso di tre anni fa. Questo è appunto ciò che *non* ha fatto lo statistico. Da un lato egli ha scelto un disegno campionario che favorisce nettamente la selezione dell'"elefante medio", ma dall'altro ha usato uno stimatore (quello di Horvitz-Thompson) del tutto inadatto a tale disegno. Infatti, lo stimatore di Horvitz-Thompson è non distorto, ma ha in questo caso una varianza elevatissima, e quindi conduce facilmente ad errori di stima molto ampi. L'errore dello statistico, in sostanza, consiste nell'aver scelto una strategia di campionamento (coppia *disegno*, *stimatore*) sostanzialmente sbagliata. Per converso, la metodologia del proprietario del circo prevede uno stimatore distorto, ma con una distorsione che, sulla base delle informazioni *a priori* note, dovrebbe esser piccola. Essendo la varianza di stima nulla, in questo caso l'errore di stima sarà presumibilmente piccolo.

14.2.7 Applicazioni a popolazioni con struttura a grappolo

Una delle applicazioni più importanti dello stimatore di Horvitz-Thompson riguarda la stima della media di popolazioni in cui le unità elementari sono raggruppate in grappoli (unità primarie). Il substrato generale è quello dei Capitoli 9, 11, e degli Esempi 12.6, 12.7.

Supponiamo che la popolazione sia divisa in M grappoli, rispettivamente di $N_1, \ldots, N_M$ unità elementari, e indichiamo al solito con $w_g = N_g/N$ il peso del grappolo g $(= 1, \ldots, M)$. Indichiamo inoltre con μ_{yg}, S^2_{yg} rispettivamente la media e la varianza corretta del grappolo g $(= 1, \ldots, M)$, e poniamo

$$z_g = M w_g \frac{T_g}{N_g}; \quad g = 1, \ldots, M$$

dove T_g è l'ammontare (il totale) del carattere $\mathcal{Y}$ nel grappolo g $(= 1, \ldots, M)$. Vale l'ovvia relazione

$$\mu_y = \frac{1}{M} \sum_{g=1}^{M} z_g.$$

I disegni campionari esposti nel Capitolo 9 (disegno a grappolo) e nel Capitolo 11 (disegno a due stadi semplici) prevedono la selezione di m grappoli mediante disegno semplice senza ripetizione. Questa scelta non è certo l'unica possibile, né è sempre la più opportuna. Vi sono parecchie situazioni, come evidenziato nella Sezione 9.4, in cui le numerosità $N_1, \ldots, N_M$ dei grappoli

sono molto differenti, ma le loro medie $\mu_{y1}, \ldots, \mu_{yM}$ possono essere ragionevolmente pensate come "molto simili". In questo caso sono i totali $T_1, \ldots, T_M$ dei grappoli ad essere molto diversi tra loro e approssimativamente proporzionali alle numerosità dei grappoli stessi. Naturalmente, per come i termini z_g e w_g sono definiti, questo equivale a dire che le z_g sono molto variabili, e approssimativamente proporzionali ai pesi w_g dei grappoli. I pesi w_g, quindi svolgono il ruolo di *misura di importanza* dei grappoli. I grappoli più "importanti" sono quelli di peso più elevato, cioè composti da un più alto numero di unità. In altre parole, i pesi dei grappoli svolgono qui il ruolo di variabile ausiliaria che misura l'importanza, la dimensione dei grappoli stessi.

In situazioni di questo tipo un'alternativa vantaggiosa al campionamento semplice dei grappoli consiste nell'usare un disegno che dia ai grappoli probabilità di inclusione proporzionale al peso w_g, fermo restando il numero m di grappoli selezionati. In termini più formali, detta $\pi_{(g)}$ la probabilità di inclusione del grappolo g $(1, \ldots, M)$, il disegno di selezione dei grappoli dovrebbe soddisfare la condizione

$$\pi_{(g)} = m\, w_g; \quad g = 1, \ldots, M. \tag{14.50}$$

Si tratta, in sostanza, di un disegno di tipo πpps (*inclusion probabilities proportional to size*). Naturalmente si assume che le (14.50) siano tutte $\leqslant 1$. Se qualcuna di esse fosse > 1, occorre porla pari a 1 e procedere come delineato nella sezione precedente.

Stimatore di Horvitz-Thompson in disegni πpps a grappolo

Nel caso di disegno a grappolo, si seleziona un campione $\mathbf{g}_m$ di m grappoli distinti, in maniera tale che le probabilità di inclusione del primo ordine dei grappoli siano del tipo (14.50). Si osservano inoltre tutte le unità elementari dei grappoli selezionati, così che i dati campionari sono $\{y_{gi};\ i = 1, \ldots, N_g;\ g \in \mathbf{g}_m\}$. In questo modo, sono anche osservate le medie μ_{yg} e le quantità z_g dei grappoli campionati. Inoltre, l'unità elementare i del grappolo g ha probabilità di inclusione del primo ordine:

$$\pi_{(g)i} = \pi_{(g)} = m\, w_g; \quad g = 1, \ldots, M.$$

Se poi si indica con $\pi_{(gg')}$ la probabilità di inclusione del secondo ordine della coppia g, g' di grappoli, la coppia i, j di unità elementari, rispettivamente dei grappoli g, g', ha probabilità di inclusione:

$$\pi_{(gg')ij} = \begin{cases} \pi_{(g)ij} = m\, w_g \ se\ g' = g \\ \pi_{(gg')} \ se\ g' \neq g \end{cases}.$$

Lo stimatore di Horvitz-Thompson della media della popolazione, μ_y,

assume in questo caso la forma

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \sum_{i=1}^{N_g} \frac{1}{\pi_{(g)i}} y_{gi} \\ &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \frac{1}{m w_g} \sum_{i=1}^{N_g} y_{gi} \\ &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \frac{1}{m w_g} N\, w_g\, \mu_{yg} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \mu_{yg} \end{aligned} \tag{14.51}$$

ossia si riduce alla media campionaria delle medie dei grappoli.

Nel caso speciale in cui i totali $T_1, \ldots, T_M$ dei grappoli sono esattamente proporzionali alle numerosità dei grappoli stessi, lo stimatore (14.51) è esattamente uguale alla media μ_y della popolazione.

Infine, è facile verificare (Esercizio 14.6) che la varianza dello stimatore (14.51) è pari a

$$V(t_{HT}) = \frac{1}{m^2} \sum_{g=1}^{M} \sum_{g'=1}^{M} \mu_{yg}\, \mu_{yg'}\, (\pi_{(gg')} - w_g w_{g'}). \tag{14.52}$$

Stimatore di Horvitz-Thompson in disegni πpps a due stadi

Un'utile alternativa al disegno campionario a grappolo precedentemente descritto consiste nell'introdurre un secondo stadio di campionamento, in cui si selezionano unità elementari dai grappoli scelti al primo stadio. Precisamente, si considera un disegno campionario del seguente tipo.

- *I stadio.* Si seleziona un campione $\mathbf{g}_m$ di m grappoli distinti, mediante un disegno che dia al generico grappolo g probabilità di inclusione del primo ordine $\pi_{(g)}$ pari alla (14.50).
- *II stadio.* Da ciascun grappolo $g \in \mathbf{g}_m$ scelto al primo stadio si seleziona, mediante disegno ssr, un campione $\mathbf{s}_g$ di n_g unità elementari. Il numero di unità elementari da selezionare da ciascun grappolo è assunto *fissato a priori*.

La logica che sostiene questo disegno, anch'esso di tipo πpps, è praticamente identica a quella del disegno a grappolo in precedenza definito. I pesi w_g svolgono il ruolo di *misura di importanza* dei grappoli, nel senso che svolgono lo stesso ruolo di una variabile ausiliaria che misura l'importanza, la dimensione dei grappoli. Questo tipo di disegno, come il precedente, dovrebbe garantire una buona efficienza allo stimatore di Horvitz-Thompson di μ_y

nel caso in cui i totali $T_1, \ldots, T_M$ dei grappoli siano molto diversi tra loro e approssimativamente proporzionali alle numerosità dei grappoli stessi.

I dati campionari sono del tipo:

$$\{y_{gi};\ i \in \mathbf{s}_g;\ g \in \mathbf{g}_m\}.$$

Poiché il secondo stadio di campionamento è semplice senza ripetizione, l'unità elementare i del grappolo g ha probabilità di inclusione del primo ordine

$$\pi_{(g)i} = \pi_{(g)} \frac{n_g}{N_g} = m\, w_g \frac{n_g}{N_g};\ \ i = 1, \ldots, N_g;\ g = 1, \ldots, M.$$

Se poi si denota con $\pi_{(gg')}$ la probabilità di inclusione del secondo ordine, al primo stadio, della coppia g, $g^{'}$ di grappoli, la coppia i, j di unità elementari, rispettivamente dei grappoli g, $g^{'}$, ha probabilità di inclusione, con ovvia simbologia, pari a:

$$\pi_{(gg')ij} = \begin{cases} \pi_{(g)ij} = m\, w_g \frac{n_g(n_g-1)}{N_g(N_g-1)} & se\ g^{'} = g \\ \pi_{(gg')} \frac{n_g}{N_g} \frac{n_{g'}}{N_{g'}} & se\ g^{'} \neq g \end{cases}.$$

La costruzione dello stimatore di Horvitz-Thompson della media della popolazione è molto semplice. Si ha:

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \sum_{i \in \mathbf{s}_g} \frac{1}{\pi_{(g)i}} y_{gi} \\ &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \frac{1}{m w_g} \sum_{i \in \mathbf{s}_g} \frac{N_g}{n_g} y_{gi} \\ &= \frac{1}{N} \sum_{g \in \mathbf{g}_m} \frac{1}{m w_g} N\, w_g \left\{ \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi} \right\} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \overline{y}_g \end{aligned} \tag{14.53}$$

essendo

$$\overline{y}_g = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi};\ \ g \in \mathbf{g}_m$$

la media campionaria del grappolo g selezionato al primo stadio.

Il calcolo della varianza dello stimatore (14.53) può essere facilmente effettuato a partire dalle probabilità di inclusione del primo e del secondo ordine dianzi calcolate. Usando anche un approccio diretto simile a quello del Capitolo 11, è ad ogni modo facile vedere (Esercizio 14.7) che:

$$V(t_{HT}) = \frac{1}{m^2} \sum_{g=1}^{M} \sum_{g'=1}^{M} \mu_{yg}\, \mu_{yg'}\, (\pi_{(gg')} - w_g w_{g'}) + \sum_{g=1}^{M} m\, w_g \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{yg}^2 \tag{14.54}$$

dove

$$S_{yg}^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2; \;\; g = 1, \ldots, M$$

è la varianza corretta del grappolo g.

*14.2.8 Efficienza dello stimatore di Horvitz-Thompson: aspetti teorici**

Si è già visto in precedenza che per garantire buone proprietà di efficienza allo stimatore di Horvitz-Thompson è necessario che il disegno di campionamento abbia ampiezza effettiva costante, e che le probabilità di inclusione del primo ordine siano, in via approssimata, proporzionali alle modalità y_i. L'obiettivo di questa sezione è quello di approfondire ed estendere questi risultati, studiando in via teorica l'efficienza dello stimatore di Horvitz-Thompson.

Poiché non esiste uno stimatore corretto di varianza uniformemente minima di μ_y, ci si concentrerà principalmente sulla proprietà di ammissibilità di t_{HT}. Dato un qualunque disegno $(\mathcal{S}, p(\cdot))$, non necessariamente ad ampiezza effettiva costante, sia $\mathcal{U}_{p,\mu_y}$ l'insieme di tutti gli stimatori corretti della media μ_y rispetto al disegno $(\mathcal{S}, p(\cdot))$. In altre parole, ogni stimatore t in $\mathcal{U}_{p,\mu_y}$ è tale che

$$E[t] = \sum_{\mathbf{s} \in \mathcal{S}} t(\boldsymbol{y}(\mathbf{s}))\, p(\mathbf{s}) = \mu_y.$$

Il primo risultato riguarda l'ammissibilità dello stimatore di Horvitz-Thompson nella classe $\mathcal{U}_{p,\mu_y}$.

Proposizione 14.7. *Qualunque sia il disegno campionario $(\mathcal{S}, p(\cdot))$, lo stimatore di Horvitz-Thompson di μ_y è ammissibile nella classe $\mathcal{U}_{p,\mu_y}$.*

Dimostrazione. Vds. Cassel *e altri* (1977), p. 55. □

La Proposizione 14.7 ci dice, in sostanza, che qualunque sia il disegno campionario adottato lo stimatore di Horvitz-Thompson della media della popolazione non può essere "uniformemente peggiore" di nessun altro stimatore corretto della media stessa. Si tratta di un buon risultato teorico, che

stabilisce una proprietà per certi aspetti minimale. I limiti della Proposizione 14.7 sono due. In primo luogo lo stimatore di Horvitz-Thompson è confrontato solo con altri stimatori corretti di μ_y. Cosa accade se si considerano anche stimatori distorti? In secondo luogo, il confronto con altri stimatori corretti è effettuato a parità di disegno campionario. Tuttavia, se si cambia lo stimatore potrebbe risultare logico cambiare anche il disegno campionario. Risulta quindi di notevole interesse effettuare il confronto non semplicemente a livello di stimatori, ma di strategie, cioè di coppie (*Disegno* , *Stimatore*). Incominciamo da quest'ultimo punto.

Indichiamo con $\mathcal{STU}_{n,\mu_y}$ la classe di tutte le strategie $(\mathcal{S}, p(\cdot)), t)$ tali che:

- il disegno $(\mathcal{S}, p(\cdot))$ è ad ampiezza effettiva costante n;
- t è un stimatore corretto rispetto al disegno $(\mathcal{S}, p(\cdot))$.

Proposizione 14.8. *Se il disegno campionario* $(\mathcal{S}, p(\cdot))$ *è ad ampiezza effettiva costante* n*, la strategia* $(\mathcal{S}, p(\cdot)), t_{HT})$ *è ammissibile nella classe* $\mathcal{STU}_{n,\mu_y}$*.*

Dimostrazione. Vds. Cassel *e altri* (1977), p. 62. □

La Proposizione 14.8 ci dice, in pratica, che l'usare una strategia formata dallo stimatore di Horvitz-Thompson e da un qualsiasi disegno ad ampiezza effettiva costante n non dà risultati uniformemente peggiori di nessun'altra strategia che utilizza un qualunque disegno campionario con ampiezza effettiva costante n e un qualunque stimatore corretto (rispetto al disegno, ovviamente).

Per quanto riguarda il confronto dello stimatore di Horvitz-Thompson della media della popolazione con altri stimatori distorti, a parità di disegno campionario, la questione è un po' più articolata. Dato un disegno $(\mathcal{S}, p(\cdot))$, indichiamo con $\mathcal{A}_{p,\mu_y}$ l'insieme di tutti gli stimatori (corretti o distorti) della media μ_y. Il risultato di base sull'ammissibilità dello stimatore di Horvitz-Thompson nella classe $\mathcal{A}_{p,\mu_y}$ è contenuto nella seguente proposizione.

Proposizione 14.9. *Se il disegno campionario* $(\mathcal{S}, p(\cdot))$ *è ad ampiezza effettiva costante, lo stimatore di Horvitz-Thompson di* μ_y *è ammissibile nella classe* $\mathcal{A}_{p,\mu_y}$*.*

Dimostrazione. Vds. Godambe e Joshi (1965). □

La Proposizione 14.9 stabilisce che *se il disegno campionario* è ad ampiezza effettiva costante, non esiste nessuno stimatore, sia esso distorto o corretto, sempre migliore di quello di Horvitz-Thompson. Si tratta, in pratica, di un'estensione della Proposizione 14.7. La condizione che il disegno campionario sia ad ampiezza effettiva costante è essenziale, ed in generale irrinunciabile. Per studiare in dettaglio questo fatto, consideriamo l'applicazione della tecnica di contrazione allo stimatore t_{HT}. In generale, la sua applicazione fornirà un nuovo stimatore

$$t_{sh*} = c^* \, t_{HT} \tag{14.55}$$

dove c^* è un numero reale definito da

$$c^* = \frac{1}{\min_{\mathbf{Y}_N \in \mathbb{R}^N} E[t_{HT}^2]/\mu_y^2} = \frac{1}{1 + \min_{\mathbf{Y}_N \in \mathbb{R}^N} V(t_{HT}^2)/\mu_y^2}. \quad (14.56)$$

Se il disegno è ad ampiezza effettiva costante, si ha

$$\min_{\mathbf{Y}_N \in \mathbb{R}^N} \frac{V(t_{HT})}{\mu_y^2} = 0 \quad (14.57)$$

in quanto, qualsiasi sia la media μ_y della popolazione, prendendo (similmente a (14.46))

$$y_i = \frac{N}{n}\,\mu_y\,\pi_i \;\; \textit{per ciascun } i = 1, \ldots, N \quad (14.58)$$

si ha $\sum_i y_i/N = \mu_y$, e

$$t_{HT}(\boldsymbol{y}(\mathbf{s})) = \mu_y \;\; \textit{per ciascun campione } \mathbf{s} \quad (14.59)$$

il che, naturalmente, equivale a $V(t_{HT}) = 0$, e quindi a $c^* = 1$. Questo, però, significa che $t_{sh*} = t_{HT}$, ossia che la tecnica di contrazione non produce alcun miglioramento dello stimatore di Horvitz-Thompson. Detto in altri termini, nel caso di disegno ad ampiezza effettiva costante lo stimatore di Horvitz-Thompson t_{HT} non è migliorabile con la tecnica di contrazione perché qualunque sia la media μ_y della popolazione esiste sempre uno speciale vettore (parametro della popolazione) $\boldsymbol{Y}_N$ che fornisce μ_y come media, e che rende lo stimatore t_{HT} identicamente uguale a μ_y, e quindi a varianza nulla.

Lo stesso ragionamento non è possibile se il disegno campionario è ad ampiezza effettiva non costante. In generale, ad eccezione di pochi casi "banali", per un generico valore di μ_y non esiste uno speciale vettore $\boldsymbol{Y}_N$ che fornisce μ_y come media, e che rende lo stimatore t_{HT} identicamente uguale a μ_y. Ma questo significa che

$$\min_{\mathbf{Y}_N \in \mathbb{R}^N} \frac{V(t_{HT})}{\mu_y^2} > 0$$

e quindi $c^* < 1$. Ne consegue che lo stimatore $t_{sh*} = c^*\, t_{HT}$ è *migliore* di t_{HT}. Il successivo esempio illustra questo punto.

Esempio 14.11 (Contrazione dello stimatore t_{HT} nel disegno scr). Supponiamo che il disegno sia semplice con ripetizione, così che le probabilità di inclusione del primo e del secondo ordine, calcolate nel Capitolo 12, sono rispettivamente pari a

$$\pi_i = 1 - \left(1 - \frac{1}{N}\right)^n = \alpha \;\; \textit{per ciascuna unità } i = 1, \ldots, N; \quad (14.60)$$

$$\begin{aligned}\pi_{ij} &= 1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n \\ &= \beta \;\; \textit{per ciascuna coppia di unità distinte}\end{aligned} \quad (14.61)$$

essendo n la numerosità campionaria. Detta come al solito $r(\mathbf{s})$ la riduzione del campione $\mathbf{s}$, e usando la simbologia introdotta in (14.60), (14.61), lo stimatore di Horvitz-Thompson di μ_y assume la forma:

$$\begin{aligned} t_{HT} &= \frac{1}{N} \sum_{i \in r(\mathbf{s})} \frac{1}{\pi_i} y_i \\ &= \frac{1}{N\,\alpha} \sum_{i \in r(\mathbf{s})} y_i \end{aligned} \tag{14.62}$$

e si ha

$$\begin{aligned} E(t_{HT}^2) &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \pi_{ij} \\ &= \frac{1}{N^2 \alpha} \sum_{i=1}^{N} y_i^2 + \frac{1}{N^2} \frac{\beta}{\alpha^2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} y_i y_j \\ &= \frac{1}{N^2} \frac{\alpha - \beta}{\alpha^2} \sum_{i=1}^{N} y_i^2 + \frac{1}{N^2} \frac{\beta}{\alpha^2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \\ &= \frac{1}{N^2} \frac{\alpha - \beta}{\alpha^2} \sum_{i=1}^{N} y_i^2 + \frac{1}{N^2} \frac{\beta}{\alpha^2} \left(\sum_{i=1}^{N} y_i \right)^2 . \end{aligned}$$

Da questa espressione si ricava che:

$$\begin{aligned} \frac{E[t_{HT}^2]}{\mu_y^2} &= \frac{\frac{1}{N^2} \frac{\alpha - \beta}{\alpha^2} \sum_{i=1}^{N} y_i^2 + \frac{1}{N^2} \frac{\beta}{\alpha^2} \left(\sum_{i=1}^{N} y_i \right)^2}{\frac{1}{N^2} \left(\sum_{i=1}^{N} y_i \right)^2} \\ &= \frac{\alpha - \beta}{\alpha^2} \frac{\sum_{i=1}^{N} y_i^2}{\left(\sum_{i=1}^{N} y_i \right)^2} + \frac{\beta}{\alpha^2}. \end{aligned} \tag{14.63}$$

Per minimizzare la (14.63) basta derivarla rispetto a $y_1, \ldots, y_N$ e annullare le derivate. Si ha:

$$\begin{aligned} \frac{\partial (E[t_{HT}^2]/\mu_y^2)}{\partial y_i} &= \frac{\alpha - \beta}{\alpha^2} \left\{ \frac{2 y_i \left(\sum_{j=1}^{N} y_j \right)^2 - 2 \left(\sum_{j=1}^{N} y_j \right) \left(\sum_{j=1}^{N} y_j^2 \right)}{\left(\sum_{j=1}^{N} y_j \right)^4} \right\} \\ &= 0 \end{aligned} \tag{14.64}$$

per ciascun $i = 1, \ldots, N$, per cui le N equazioni (14.64) hanno come soluzione

$$y_1 = y_2 = \cdots = y_N$$

che ovviamente equivale a

$$y_i = \mu_y \;\; per \; ciascun \; i = 1, \ldots, N. \tag{14.65}$$

Dalla (14.65) si desume la relazione

$$\min_{\mathbf{Y}_N \in \mathbb{R}^N} \frac{E[t_{HT}^2]}{\mu_y^2} = \frac{N\,\alpha + N(N-1)\,\beta}{(N\,\alpha)^2}$$

da cui, tenendo anche conto che

$$N\,\alpha = \sum_{i=1}^{N} \pi_i = \overline{\nu}$$

$$N(N-1)\,\beta = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = V(\nu(\mathbf{s})) + \overline{\nu}(\overline{\nu} - 1)$$

essendo $\nu(\mathbf{s})$ l'ampiezza effettiva del campione $\mathbf{s}$, $\overline{\nu}$ la sua media e $V(\nu(\mathbf{s}))$ la sua varianza, si ottiene infine

$$\min_{\mathbf{Y}_N \in \mathbb{R}^N} \frac{E[t_{HT}^2]}{\mu_y^2} = 1 + \frac{V(\nu(\mathbf{s}))}{\overline{\nu}^2} \tag{14.66}$$

e quindi

$$c^* = \frac{1}{1 + \frac{V(\nu(\mathbf{s}))}{\overline{\nu}^2}}. \tag{14.67}$$

Dalla (14.66) si vede con facilità che $c^* < 1$. Pertanto, lo stimatore

$$t_{sh*} = c^*\, t_{HT} \tag{14.68}$$

con c^* dato dalla (14.67), è migliore dello stimatore di Horvitz-Thompson t_{HT}.

È interessante confrontare questo risultato con quello dell'Esempio 13.10, dove si è mostrato che la media campionaria delle unità distinte, $\overline{y}_{r(\mathbf{s})}$, non è migliorabile con la tecnica di contrazione. Malgrado questo sia un punto a favore di $\overline{y}_{r(\mathbf{s})}$, non si può affermare che esso sia migliore dello stimatore di Horvitz-Thompson (14.62), in quanto la Proposizione 14.7 stabilisce che nessuno stimatore corretto di μ_y può essere migliore di quello di Horvitz-Thompson. □

Tra i casi "banali" di disegni campionari ad ampiezza effettiva non costante ma tali che lo stimatore di Horvitz-Thompson non è migliorabile con la tecnica di contrazione ve ne sono alcuni molto importanti. Di seguito sono forniti alcuni esempi in proposito.

Esempio 14.12 (Disegno a grappolo "semplice"). Consideriamo ancora l'Esempio 14.6, in cui si è costruito lo stimatore di Horvitz-Thompson per un disegno a grappolo "semplice" (ossia con uguali probabilità di selezione dei grappoli). Usando la stessa notazione dell'Esempio 14.6, lo stimatore di Horvitz-Thompson assume la forma

$$t_{HT} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \mu_{yg}. \tag{14.69}$$

Per verificare se t_{HT} è migliorabile con la tecnica di contrazione, bisogna calcolare:

$$\min \frac{E[t_{HT}^2]}{\mu_y^2} = \min \frac{V(t_{HT}) + 1}{\mu_y^2} = 1 + \min \frac{V(t_{HT})}{\mu_y^2}. \tag{14.70}$$

Ora, dato un qualsiasi valore di μ_y, se si considerano i valori

$$y_{gi} = \frac{\mu_y}{M\, w_g}, \;\; i = 1, \ldots, N_g; \; g = 1, \ldots M$$

si ha che la loro media è μ_y, e che $\mu_{yg} = \mu_y/(M\, w_g)$ per ciascun grappolo $g = 1, \ldots, M$, per cui è anche

$$t_{HT} = \mu_y \;\; \textit{per ciascun campione } \mathbf{g}_m.$$

Ma ciò significa che per questa "speciale popolazione" si ha $V(t_{HT}) = 0$, da cui, usando la (14.70),

$$\min \frac{E[t_{HT}^2]}{\mu_y^2} = 1$$

e $c^* = 1$. In questo caso lo stimatore di Horvitz-Thompson non è migliorabile con la tecnica di contrazione. Lo stesso tipo di risultato vale anche nel caso di disegno a grappolo πpps, in cui lo stimatore di Horvitz-Thompson della media assume la forma (14.51) (vds. Esercizio 14.8). □

Esempio 14.13 (Disegno a due stadi semplici). Consideriamo ancora l'Esempio 14.7, in cui si è costruito lo stimatore di Horvitz-Thompson per un disegno a due stadi semplici. Con la notazione in precedenza usata, lo stimatore di Horvitz-Thompson assume la forma

$$t_{HT} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} M w_g \overline{y}_g. \tag{14.71}$$

È facile vedere che qualunque sia la media μ_y la stessa "popolazione speciale" dell'Esempio 14.12 possiede μ_y come media, e rende nulla la varianza dello stimatore di Horvitz-Thompson. Ma questo significa che $c^* = 1$, per cui lo stimatore di Horvitz-Thompson non è migliorabile con la tecnica di contrazione. Lo stesso tipo di risultato vale anche nel caso di disegno a due stadi πpps, in cui lo stimatore di Horvitz-Thompson della media assume la forma (14.53) (vds. Esercizio 14.9). □

14.3 Variazioni sul tema: stimatore alle differenze generalizzate

Sia $\mathbf{s}$ un campione di numerosità n ottenuto mediante un prefissato disegno campionario $(\mathcal{S}, p(\cdot))$. In corrispondenza di ogni unità i della popolazione, sia poi y_i^0 $(i = 1, \ldots, N)$ un arbitrario numero reale (noto). In linea di principio, y_i^0 dovrebbe essere, per quanto possibile, una "approssimazione" dell'incognita modalità y_i del carattere di interesse $\mathcal{Y}$. Indichiamo con

$$\mu_0 = \frac{1}{N} \sum_{i=1}^{N} y_i^0$$

la media dei numeri $y_1^0, \ldots, y_N^0$.

L'idea alla base dello stimatore alle differenze generalizzate è di utilizzare i valori $y_1^0, \ldots, y_N^0$ per costruire una stima delle media μ_y della popolazione. Il punto di partenza consiste nell'osservare che questa può essere scritta come

$$\begin{aligned} \mu_y &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i^0 + \frac{1}{N} \sum_{i=1}^{N} (y_i - y_i^0) \\ &= \mu_0 + \frac{1}{N} \sum_{i=1}^{N} e_i \\ &= \mu_0 + \mu_e \end{aligned} \tag{14.72}$$

dove si è posto

$$\begin{aligned} e_i &= y_i - y_i^0, \quad i = 1, \ldots, N; \\ \mu_e &= \frac{1}{N} \sum_{i=1}^{N} e_i. \end{aligned}$$

Chiaramente, la media μ_0 dei valori y_i^0 è nota, mentre la media μ_e delle differenze $e_i = y_i - y_i^0$ è incognita. L'idea alla base dello stimatore alle differenze generalizzate è quella di stimare μ_e utilizzando uno stimatore di Horvitz-Thompson. Formalmente

$$\begin{aligned} t_{gd} &= \mu_0 + \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} (y_i - y_i^0) \\ &= \mu_0 + \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} (y_i - y_i^0)\, \delta(i;\, \mathbf{s})\,. \end{aligned} \tag{14.73}$$

Ribadiamo che il primo termine della (14.73) è una quantità fissa, costante e indipendente dal campione, mentre il secondo termine è uno stimatore di Horvitz-Thompson applicato alle differenze $y_i - y_i^0$. In sostanza, quindi, a meno della costante additiva μ_0, lo stimatore alle differenze generalizzate (14.73) è sostanzialmente uno stimatore di tipo Horvitz-Thompson. In simboli:

$$t_{gd} = \mu_0 + t_{HTe} \tag{14.74}$$

dove

$$\begin{aligned} t_{HTe} &= \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} e_i \\ &= \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} (y_i - y_i^0) \end{aligned} \tag{14.75}$$

è lo stimatore di Horvitz-Thompson di μ_e.

A causa della (14.75), lo stimatore (14.73) eredita le sue proprietà da quelle dello stimatore di Horvitz-Thompson.

Le proprietà dello stimatore (14.73) sono riassunte nella seguente proposizione.

Proposizione 14.10. *Lo stimatore alle differenze generalizzate t_{gd} è uno stimatore corretto della media della popolazione:*

$$E[t_{gd}] = \mu_y. \tag{14.76}$$

La varianza di t_{gd} è pari a

$$V(t_{gd}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{y_i - y_i^0}{\pi_i} \frac{y_j - y_j^0}{\pi_j} (\pi_{ij} - \pi_i \pi_j). \tag{14.77}$$

Dimostrazione. È sufficiente usare la (14.75) e le ben note proprietà dello stimatore di Horvitz-Thompson. □

Per quanto riguarda la stima della varianza di t_{gd}, basta ovviamente applicare i risultati già visti per la stima della varianza dello stimatore di Horvitz-Thompson.

Dalla (14.77) è evidente che lo stimatore alle differenze generalizzate è tanto più efficiente quanto più le differenze e_i risultano prossime allo zero per ogni unità della popolazione. Quindi lo scopo è usare valori y_i^0 tali che le differenze $e_i = y_i - y_i^0$ risultino "piccole". Un'idea molto semplice consiste nell'utilizzare valori y_i^0 proporzionali ai valori x_i ($i = 1, \ldots, N$) di una variabile ausiliaria $\mathcal{X}$ nota per tutte le unità della popolazione. Formalmente:

$$y_i^0 = c\, x_i, \quad i = 1, \ldots, N$$

dove c è una costante di proporzionalità. In questo caso lo stimatore alle differenze generalizzate può essere scritto come

$$\begin{aligned} t_{gd} &= \frac{1}{N}\sum_{i=1}^{N} y_i^0 + \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{(y_i - y_i^0)}{\pi_i} \\ &= \frac{1}{N}\sum_{i=1}^{N} cx_i + \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{(y_i - cx_i)}{\pi_i} \\ &= c\,\mu_x + \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{y_i}{\pi_i} - c\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{x_i}{\pi_i} \\ &= \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{y_i}{\pi_i} - c\left(\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{x_i}{\pi_i} - \mu_x\right) \\ &= t_{HTy} - c(t_{HTx} - \mu_x) \end{aligned} \tag{14.78}$$

dove

$$t_{HTy} = \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{y_i}{\pi_i}, \quad t_{HTx} = \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{x_i}{\pi_i}$$

sono gli stimatori di Horvitz-Thompson rispettivamente di μ_y e μ_x.

Sulla base della (14.78) lo stimatore alle differenze generalizzate è dato dalla differenza tra stimatore di Horvitz-Thompson t_{HTy} della media della popolazione e un termine di aggiustamento pari a $c(t_{HTx} - \mu_x)$.

È facile infine provare (Esercizio 14.13) che sotto un disegno semplice senza ripetizione di ampiezza n si ottiene lo stimatore alle differenze della Sezione 5.2.

14.4 Vecchie glorie un po' in disarmo: lo stimatore di Hansen-Hurwitz

Data una popolazione finita di N unità, si consideri un disegno campionario (ordinato e con ripetizioni) di tipo *ppswr* di ampiezza n, introdotto nella Sezione 12.4.1. Ad ogni unità i della popolazione è associato *a priori* un numero p_i ($i = 1, \ldots, N$), che ne esprime in qualche modo l'"importanza". Lo spazio $\mathcal{S}$ dei campioni di unità è l'insieme di tutte le n-ple ordinate (disposizioni con ripetizioni di classe n) del tipo $(i_1, i_2, \ldots, i_n)$, in cui i_1 è la *prima* unità del campione, i_2 è la *seconda* unità del campione, e così via. Inoltre, $i_1, i_2, \ldots, i_n$ possono essere unità qualsiasi della popolazione, senza alcuna limitazione o vincolo. In modo più formale, questo significa, come già detto, che

$$\mathcal{S} = \underbrace{I_N \times I_N \times \cdots \times I_N}_{n \text{ volte}} = I_N^n.$$

Per quanto riguarda le probabilità dei campioni, se $\mathbf{s} = (i_1, i_2, \ldots, i_n)$ si ha

$$p(\mathbf{s}) = p_{i_1} p_{i_2} \cdots p_{i_n}.$$

In termini un po' più intuitivi, è come se si effettuassero n "prove"; nella prima prova si seleziona i_1, la prima unità del campione, nella seconda prova si seleziona i_2, la seconda unità del campione, e così via. I risultati delle diverse prove sono indipendenti, hanno identica distribuzione, e sono tali che, per ciascuna prova $k = 1, \ldots, n$,

$$Pr(i_k = i) = p_i \quad \textit{per ogni unità } i = 1, \ldots, N. \tag{14.79}$$

Indichiamo poi con y_{i_k} la modalità dell'unità selezionata nella prova k $(= 1, \ldots, n)$. Come conseguenza della (14.79), anche le variabili aleatorie $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$ sono indipendenti e hanno la stessa distribuzione di probabilità. Ne consegue che sono indipendenti e identicamente distribuite anche le variabili aleatorie

$$\frac{y_{i_1}}{N p_{i_1}}, \frac{y_{i_2}}{N p_{i_2}}, \ldots, \frac{y_{i_n}}{N p_{i_n}}.$$

Le distribuzioni di probabilità delle variabili dianzi introdotte sono descritte nella Tabella 14.1.

Tabella 14.1 Distribuzione delle variabili aleatorie nel disegno *ppswr* per la prova $k$ma

i_k	*Probabilità*	y_{i_k}	$\frac{y_{i_k}}{N p_{i_k}}$
1	p_1	y_1	$\frac{y_1}{N p_1}$
2	p_2	y_2	$\frac{y_2}{N p_2}$
...	...	...	...
i	p_i	y_i	$\frac{y_i}{N p_i}$
...	...	...	...
N	p_N	y_N	$\frac{y_N}{N p_N}$

In particolare, dalla Tabella 14.1 si ricava facilmente (Esercizio 14.14) che

$$\begin{aligned} E\left[\frac{y_{i_k}}{N p_{i_k}}\right] &= \frac{y_1}{N p_1} p_1 + \frac{y_2}{N p_2} p_2 + \cdots + \frac{y_N}{N p_N} p_N \\ &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ &= \mu_y \quad \textit{per ciascun } k = 1, \ldots, n; \end{aligned} \tag{14.80}$$

$$V\left(\frac{y_{i_k}}{Np_{i_k}}\right) = \sum_{i=1}^{N}\left(\frac{y_i}{Np_i} - \mu_y\right)^2 p_i \tag{14.81}$$

$$= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 p_i\, p_j$$

$$\text{per ciascun } k = 1, \ldots, n. \tag{14.82}$$

Lo *stimatore di Hansen-Hurwitz* della media μ_y della popolazione è definito come

$$t_{HH} = \frac{1}{n}\sum_{i\in\mathbf{s}}\frac{y_i}{Np_i} \tag{14.83}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\frac{y_{i_k}}{Np_{i_k}}. \tag{14.84}$$

Le (14.83) e (14.84) sono perfettamente equivalenti. In particolare, dalla (14.83) risulta evidente che t_{HH} non è altro che la media campionaria delle quantità $y_i/(Np_i)$. La (14.84) mette inoltre in evidenza che le quantità $y_i/(Np_i)$ sono realizzazioni di variabili aleatorie indipendenti e identicamente distribuite. Quest'osservazione si rivela utile per lo studio delle proprietà dello stimatore di Hansen-Hurwitz, riportate nella proposizione successiva.

Proposizione 14.11. *Lo stimatore di Hansen-Hurwitz è corretto*

$$E[t_{HH}] = \mu_y \tag{14.85}$$

e la sua varianza è pari a

$$V(t_{HH}) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{Np_i} - \mu_y\right)^2 p_i. \tag{14.86}$$

Inoltre, uno stimatore corretto di $V(t_{HH})$ è il seguente

$$\widehat{V}(t_{HH}) = \frac{1}{n-1}\sum_{i\in\mathbf{s}}\left(\frac{y_i}{Np_i} - t_{HH}\right)^2. \tag{14.87}$$

Dimostrazione. Basta tener conto che le variabili aleatorie $y_{i_1}/(Np_{i_1}), \ldots, y_{i_n}/(Np_{i_1})$ sono indipendenti e identicamente distribuite, con media e varianza rispettivamente pari a (14.80) e (14.81), e che t_{HH} è la loro media campionaria. Si ha:

$$E[t_{HH}] = \frac{1}{n}\sum_{k=1}^{n} E\left[\frac{y_{i_k}}{Np_{i_k}}\right]$$

$$= \mu_y$$

$$V(t_{HH}) = \frac{1}{n^2}\sum_{k=1}^{n} V\left(\frac{y_{i_k}}{Np_{i_k}}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{Np_i} - \mu_y\right)^2 p_i.$$

Per la (14.87), infine, basta tener conto che

$$\widehat{V}_{HH} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{y_{i_k}}{Np_{i_k}} - t_{HH}\right)^2$$

per cui $\widehat{V}(t_{HH})$ è null'altro che la varianza campionaria corretta per un campione composto da n variabili aleatorie indipendenti e identicamente distribuite. □

Esempio 14.14. Nel caso speciale in cui

$$p_1 = p_2 = \cdots = p_N = \frac{1}{N}$$

il disegno *ppswr* si riduce al classico disegno campionario semplice con ripetizione (scr). Essendo $Np_i = 1$, lo stimatore di Hansen-Hurwitz si riduce alla media campionaria

$$t_{HH} = \tfrac{1}{n}\textstyle\sum_{i\in\mathbf{s}} y_i = \overline{y}_{\mathbf{s}}.$$

La Proposizione 14.11 mette in evidenza (come già visto nel Capitolo 3) che $\overline{y}_{\mathbf{s}}$ (usata con il disegno scr) è uno stimatore corretto della media μ_y della popolazione, e che la sua varianza è pari a

$$V(t_{HH}) = V(\overline{y}_{\mathbf{s}}) = \tfrac{1}{n}\sigma_y^2$$

essendo $\sigma_y^2 = \sum_{i=1}^{N}(y_i - \mu_y)^2/N$ la varianza della popolazione. □

L'espressione della varianza dello stimatore di Hansen-Hurwitz mette in evidenza un fatto importante, utile per la scelta delle probabilità p_i di selezione delle unità. Infatti, dalla (14.86) e dalla (14.82) si vede facilmente che

$$V(t_{HH}) = \frac{1}{n}\left\{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2 p_i\, p_j\right\} \tag{14.88}$$

e la (14.88) è pari a 0 se le probabilità di selezione p_i soddisfano la relazione

$$p_i = \frac{y_i}{N\mu_y} \quad \textit{per ciascun } i = 1, \ldots, N. \tag{14.89}$$

La (14.89) mette in evidenza un fatto importante: *lo stimatore di Hansen-Hurwitz ha massima efficienza quando le probabilità di selezione p_i sono proporzionali alle modalità y_i*. Si tratta di un risultato simile a quello visto nella Sezione 14.2.6 per lo stimatore di Horvitz-Thompson, benché in quel caso fossero coinvolte le probabilità di inclusione π_i, che sono diverse dalle p_i.

Naturalmente, scegliere le probabilità di selezione (14.89) richiede la conoscenza delle modalità y_i del carattere di interesse, che sono incognite. Un ripiego ragionevole consiste nello scegliere probabilità p_i proporzionali ai valori di una variabile ausiliaria $\mathcal{X}$. In simboli:

$$p_i = \frac{x_i}{N\mu_x} \quad \textit{per ciascun } i = 1, \ldots, N. \tag{14.90}$$

Tale scelta dà risultati tanto migliori quanto più le modalità x_i della variabile ausiliaria sono "vicine" ad una situazione di proporzionalità rispetto alle modalità y_i della variabile di interesse. Al limite, se fosse $y_i/x_i = \textit{costante}$, le (14.90) coinciderebbero con le (14.89)

Malgrado le molte proprietà positive, lo stimatore di Hansen-Hurwitz soffre di un serio inconveniente che ne ha limitato fortemente l'uso: *dipende dalle ripetizioni delle unità campionarie, e quindi è migliorabile per il Teorema di Rao-Blackwell*. Detta $\boldsymbol{y}(r(\mathbf{s}))$ la riduzione dei dati campionari, lo stimatore

$$t^* = E[t_{HH} \,|\boldsymbol{y}(r(\mathbf{s}))] \tag{14.91}$$

è più efficiente di t_{HH}. Purtroppo, ad eccezione del caso $p_1 = \cdots = p_N = 1/N$, per il quale si rinvia all'Esempio 13.5, non è facile costruire lo stimatore (14.91). Il caso di campioni di numerosità $n = 3$ è trattato nell'Esercizio 13.6.

Una delle applicazioni più interessanti dello stimatore di Hansen-Hurwitz è ai disegni a grappoli e a due stadi, in cui la selezione dei grappoli avviene mediate disegno *ppswr* con probabilità di selezione pari ai pesi dei grappoli stessi. Ciò è illustrato nei successivi esempi.

Esempio 14.15 (Disegno a grappolo *ppswr*). Si consideri una popolazione suddivisa in M grappoli, rispettivamente di $N_1, \ldots, N_M$ unità elementari. Siano inoltre, come al solito, $w_g = N_g/N$ e μ_{yg} rispettivamente il peso e la media del grappolo g $(= 1, \ldots, M)$. Un processo di selezione dei grappoli molto semplice consiste nel selezionare m grappoli mediante disegno *ppswr* con probabilità di selezione dei grappoli $p_g = w_g$, e nell'osservare tutte le unità elementari dei grappoli selezionati. Detto $\mathbf{g}_m$ il campione di grappoli selezionato, e tenendo conto che $\mu_y = \sum_g M\, w_g\, \mu_{yg}/M$, lo stimatore di Hansen-Hurwitz di μ_y assume la forma:

$$\begin{aligned} t_{HHgr} &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \frac{M\, w_g\, \mu_{yg}}{M\, w_g} \\ &= \frac{1}{m} \sum_{g \in \mathbf{g}_m} \mu_{yg}. \end{aligned} \tag{14.92}$$

È facile verificare che lo stimatore t_{HHgr} è corretto, con varianza

$$V(t_{HHgr}) = \frac{1}{2m} \sum_{g=1}^{M} \sum_{g'=1}^{M} \left(\mu_{yg} - \mu_{yg'} \right)^2 . \qquad \square$$

Esempio 14.16 (Disegno a due stadi *ppswr*). Una facile variante del disegno campionario a grappolo *ppswr* consiste nell'introdurre un secondo stadio di campionamento, in cui si selezionano unità elementari dai grappoli scelti al primo stadio. Precisamente, si considera un disegno campionario del seguente tipo.

- *I stadio.* Si seleziona un campione $\mathbf{g}_m$ di m grappoli (non necessariamente distinti), mediante un disegno *ppswr* che dia al generico grappolo g probabilità di selezione $p_g = w_g$.
- *II stadio.* Da ciascun grappolo $g \in \mathbf{g}_m$ scelto al primo stadio si seleziona, mediante disegno ssr, un campione $\mathbf{s}_g$ di n_g unità elementari. Il numero di unità elementari da selezionare da ciascun grappolo è assunto *fissato a priori*.

La costruzione di uno stimatore di tipo Hansen-Hurwitz della media della popolazione è molto semplice. Si ha:

$$t_{HH2st} = \frac{1}{m} \sum_{g \in \mathbf{g}_m} \overline{y}_g \tag{14.93}$$

dove

$$\overline{y}_g = \frac{1}{n_g} \sum_{i \in \mathbf{s}_g} y_{gi}; \quad g \in \mathbf{g}_m$$

è la media campionaria del grappolo g selezionato al primo stadio.

Dalla relazione (immediata da verificare)

$$E[t_{HH2st} \,|\mathbf{g}_m] = t_{HHgr}$$

con t_{HHgr} dato dalla (14.92), si vede subito che t_{HH2st} è corretto. La sua varianza, facilmente calcolabile con un approccio diretto simile a quello del Capitolo 11 (Esercizio 14.15) è pari a:

$$V(t_{HH2st}) = \frac{1}{2m} \sum_{g=1}^{M} \sum_{g'=1}^{M} \left(\mu_{yg} - \mu_{yg'} \right)^2 + \sum_{g=1}^{M} \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_{yg}^2 \, w_g \tag{14.94}$$

dove

$$S_{yg}^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (y_{gi} - \mu_{yg})^2; \quad g = 1, \ldots, M$$

è la varianza corretta del grappolo g. $\square$

14.5 Largo ai giovani: qualche idea di base sugli stimatori di tipo calibrazione*

14.5.1 Calibrazione con una variabile ausiliaria

Come visto a proposito dello stimatore di Horvitz-Thompson, una scelta "ragionevole" delle probabilità di inclusione π_i consiste nel prenderle proporzionali ai valori di una variabile ausiliaria, il più possibile "vicina", a sua volta, ad un legame di approssimata proporzionalità con la variabile di interesse. In questo modo, la variabile ausiliaria viene usata sia per costruire il disegno campionario di selezione delle unità, sia per costruire lo stimatore (di Horvitz-Thompson) di μ_y. Tuttavia, questo non è l'unico modo di utilizzare variabili ausiliarie per stimare la media della variabile di interesse. Ad esempio, anche gli stimatori per quoziente e per regressione usano una variabile ausiliaria (indipendentemente dal tipo di disegno campionario adottato, che può o meno essere di tipo ssr).

Gli stimatori di tipo calibrazione sono stati introdotti abbastanza di recente come tentativo di tener conto esplicitamente, in fase di stima, delle informazioni derivanti dalla conoscenza di una o più variabili ausiliarie note. In questa sezione ci si limiterà soltanto a fornire alcune idee di base sugli stimatori per calibrazione. Una trattazione più dettagliata, basata su un *modello statistico* che espliciti in che modo la variabile di interesse $\mathcal{Y}$ dipende dalle variabili ausiliarie note, sarà fornita nella parte riguardante i modelli di superpopolazione.

L'idea di base dell'approccio basato sulla calibrazione può essere esposta in modo molto semplice. Supponiamo di disporre di informazioni ausiliarie consistenti nella conoscenza di un carattere ausiliario $\mathcal{X}$, correlato con il carattere di interesse $\mathcal{Y}$, e di cui:

- è nota la media $\mu_x = \sum_{i=1}^{N} x_i/N$;
- sono noti i valori x_i assunti in corrispondenza delle unità campionarie.

Si osservi che se i valori x_i sono noti per tutte le unità della popolazione (come in genere accade in pratica) le due assunzioni precedenti sono soddisfatte.

Se $\mathbf{s}$ è il campione di unità selezionate, sia

$$t_{HTy} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} y_i$$

lo stimatore di Horvitz-Thompson di μ_y, e sia

$$t_{HTx} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} x_i$$

il corrispondente stimatore di Horvitz-Thompson di μ_x.

Se t_{HTx} è "lontano" da μ_x, e se $\mathcal{Y}$ è correlato abbastanza fortemente con $\mathcal{X}$, è ragionevole attendersi che anche t_{HTy} sia "lontano" da μ_y. L'idea di base è allora quella di stimare μ_y con uno stimatore lineare

$$t_{ly} = \frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, y_i \tag{14.95}$$

che soddisfi i seguenti due requisiti.

1. La stima della media della variabile $\mathcal{X}$ ottenuta applicando i pesi finali $c_{i\mathbf{s}}$ ai dati $\{x_i;\ i \in \mathbf{s}\}$ deve uguagliare la media della popolazione μ_x. Tale vincolo è denominato vincolo di calibrazione. Formalmente

$$t_{lx} = \frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, x_i = \mu_x. \tag{14.96}$$

2. I pesi finali $c_{i\mathbf{s}}$ sono il più possibile "vicini" a quelli base $1/\pi_i$ dello stimatore di Horvitz-Thompson determinati sulla base del disegno di campionamento.

Naturalmente, il requisito 2 richiede di definire una distanza tra i pesi $c_{i\mathbf{s}}$ e le $1/\pi_i$. La distanza adottata è la seguente:

$$\sum_{i \in \mathbf{s}} \frac{(c_{i\mathbf{s}} - \frac{1}{\pi_i})^2}{\frac{q_i}{\pi_i}} \tag{14.97}$$

dove q_i sono numeri positivi arbitrari.

La determinazione dei pesi finali $c_{i\mathbf{s}}$ che definiscono lo stimatore (14.95) richiede quindi la soluzione del seguente problema di ottimo vincolato:

$$\begin{cases} \textit{minimizzare}: \ \sum_{i \in \mathbf{s}} \frac{(c_{i\mathbf{s}} - \frac{1}{\pi_i})^2}{\frac{q_i}{\pi_i}} \\ \textit{con il vincolo}: \ \frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, x_i = \mu_x. \end{cases} \tag{14.98}$$

Proposizione 14.12. *La soluzione del problema di minimo vincolato* (14.98) *è costituita dai pesi:*

$$c_{i\mathbf{s}} = \frac{1}{\pi_i} - \frac{t_{HTx} - \mu_x}{\frac{1}{N} \sum_{i \in \mathbf{s}} \frac{q_i x_i^2}{\pi_i}} \frac{q_i x_i}{\pi_i}. \tag{14.99}$$

Dimostrazione. Il problema di minimo vincolato (14.98) si risolve con il metodo dei moltiplicatori di Lagrange. La funzione Lagrangiana assume la forma

$$\mathcal{L} = \sum_{i \in \mathbf{s}} \frac{(c_{i\mathbf{s}} - \frac{1}{\pi_i})^2}{\frac{q_i}{\pi_i}} - 2\lambda \left(\sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, x_i - N\mu_x \right) \tag{14.100}$$

dove λ è il moltiplicatore di Lagrange. Derivando la (14.100) rispetto a $c_{i\mathbf{s}}$ e annullando tali derivate, si ha

$$\frac{\partial \mathcal{L}}{\partial c_{i\mathbf{s}}} = 2\frac{c_{i\mathbf{s}} - \frac{1}{\pi_i}}{\frac{q_i}{\pi_i}} - 2\lambda\, x_i = 0$$

che equivale a

$$c_{i\mathbf{s}} = \frac{1}{\pi_i} + \lambda\, \frac{q_i x_i}{\pi_i}. \tag{14.101}$$

Moltiplicando ambo i membri della (14.101) per x_i, si ha

$$c_{i\mathbf{s}}\, x_i = \frac{x_i}{\pi_i} + \lambda\, \frac{q_i x_i^2}{\pi_i}. \tag{14.102}$$

Usando infine il vincolo di calibrazione (14.96), dalla (14.102) si ottiene

$$\mu_x = \frac{1}{N}\sum_{i\in\mathbf{s}} c_{i\mathbf{s}}\, x_i = t_{HTx} + \frac{\lambda}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i^2}{\pi_i}$$

da cui

$$\lambda = \frac{\mu_x - t_{HTx}}{\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i^2}{\pi_i}}$$

e quindi la (14.99). Si osservi infine che il peso finale può essere espresso come prodotto tra il peso base e un fattore correttivo. □

L'uso dei pesi (14.99) porta allo *stimatore per calibrazione* di μ_y, definito come

$$\begin{aligned} t_{cal} &= \frac{1}{N}\sum_{i\in\mathbf{s}} \left\{ \frac{1}{\pi_i} - \frac{t_{HTx} - \mu_x}{\frac{1}{N}\sum_{j\in\mathbf{s}} \frac{q_j x_j^2}{\pi_j}} \frac{q_i x_i}{\pi_i} \right\} y_i \\ &= \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{1}{\pi_i}\, y_i - \frac{\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i}{\pi_i} y_i}{\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i^2}{\pi_i}} (t_{HTx} - \mu_x) \\ &= t_{HTy} - \frac{\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i}{\pi_i} y_i}{\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i x_i^2}{\pi_i}} (t_{HTx} - \mu_x). \end{aligned} \tag{14.103}$$

Intuitivamente, lo stimatore (14.103) "aggiusta", "calibra" lo stimatore di Horvitz-Thompson t_{HTy} con un termine che dipende dalla differenza tra t_{HTx} e μ_x, ossia dalla "distanza" tra lo stimatore di Horvitz-Thompson di μ_x, t_{HTx} e l' effettiva μ_x. Quanto più grande è (in positivo o in negativo) la

differenza tra t_{HTx} e μ_x, tanto più grande sarà l'aggiustamento. Il "fattore di aggiustamento" è rappresentato dal rapporto

$$\frac{\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{q_i x_i}{\pi_i}y_i}{\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{q_i x_i^2}{\pi_i}}$$

in cui il numeratore è lo stimatore di Horvitz-Thompson di $\sum_{i=1}^{N} q_i\, x_i\, y_i/N$, mentre il denominatore è lo stimatore di Horvitz-Thompson di $\sum_{i=1}^{N} q_i\, x_i^2/N$.

Esempio 14.17. Si supponga che $q_i = 1/x_i$ per ciascuna unità della popolazione. Si ha allora

$$\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{q_i\, x_i}{\pi_i}y_i = \frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}\, y_i = t_{HTy}$$
$$\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{q_i\, x_i^2}{\pi_i} = \frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}\, x_i = t_{HTx}$$

per cui lo stimatore per calibrazione si riduce a

$$\begin{aligned} t_{cal} &= t_{HTy} - \frac{t_{HTy}}{t_{HTx}}\,(t_{HTx} - \mu_x) \\ &= \frac{t_{HTy}}{t_{HTx}}\,\mu_x. \end{aligned} \tag{14.104}$$

Si tratta, in sostanza, di una versione generalizzata del classico stimatore per quoziente. Se poi il disegno campionario è semplice senza ripetizione lo stimatore di Horvitz-Thompson si riduce alla media campionaria, e lo stimatore (14.104) diventa proprio lo stimatore per quoziente. □

Esempio 14.18. Consideriamo ancora l'Esempio 14.17, e assumiamo che la variabile ausiliaria $\mathcal{X}$ sia tale che $x_i = 1$ per ciascuna unità della popolazione, così che è anche $\mu_x = 1$. Di fatto, non si tratta di una "vera" informazione ausiliaria. Il vincolo di calibrazione (14.96) assume la forma:

$$\frac{1}{N}\sum_{i\in\mathbf{s}} c_{i\mathbf{s}} = 1$$

che equivale a

$$\sum_{i\in\mathbf{s}} c_{i\mathbf{s}} = N. \tag{14.105}$$

Se si interpreta il peso $c_{i\mathbf{s}}$ come il numero di unità della popolazione rappresentate dall'unità i del campione $\mathbf{s}$, allora la somma $\sum_{i\in\mathbf{s}} c_{i\mathbf{s}}$ rappresenta il numero totale di unità della popolazione rappresentate dalle unità del campione $\mathbf{s}$. Il vincolo (14.105) ci dice che il *numero totale di unità rappresentate*

da quelle campionarie deve essere eguale all'effettivo numero di unità della popolazione.

Se si assume anche $q_i = 1$ per tutte le unità della popolazione, lo stimatore di calibrazione (14.104) diviene

$$t_{cal} = \frac{\sum_{i \in \mathbf{s}} \frac{1}{\pi_i} y_i}{\sum_{i \in \mathbf{s}} \frac{1}{\pi_i}}. \tag{14.106}$$

Si tratta di uno stimatore di tipo quoziente generalizzato introdotto da Hájek (1971), in cui i pesi dello stimatore hanno la forma:

$$c_{i\mathbf{s}} = N \left(\frac{1}{\pi_i} \Big/ \sum_{j \in \mathbf{s}} \frac{1}{\pi_j} \right) \quad \textit{per ciascun } i \in \mathbf{s}. \qquad \square$$

14.5.2 Calibrazione con più variabili ausiliarie

Le idee della precedente sezione possono facilmente essere estese al caso di più variabili ausiliarie. Supponiamo date p variabili ausiliarie $\mathcal{X}_1, \ldots, \mathcal{X}_p$, e indichiamo con x_{ik} il valore che la variabile $\mathcal{X}_k$ assume in corrispondenza dell'unità i ($i = 1, \ldots, N$; $k = 1, \ldots, p$). Indichiamo inoltre con

$$\mu_{x_k} = \frac{1}{N} \sum_{i=1}^{N} x_{ik}; \quad k = 1, \ldots, p \tag{14.107}$$

la media della variabile $\mathcal{X}_k$ ($k = 1, \ldots, p$).

Nel seguito si assumerà che le p medie (14.107) siano note, e che i valori x_{ik} siano noti per tutte le unità campionarie. Non è necessario, a questo stadio, assumere i valori x_{ik} noti per tutte le unità della popolazione.

Le idee che hanno portato allo stimatore per calibrazione nella sezione precedente rimangono sostanzialmente invariate. Il problema è quello di costruire uno stimatore lineare di μ_y del tipo (14.95), in modo tale che (i) sia minima la distanza (14.97), e (ii) siano soddisfatti i p *vincoli di calibrazione*

$$\frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, x_{ik} = \mu_{x_k}; \quad k = 1, \ldots, p. \tag{14.108}$$

La determinazione dei pesi $c_{i\mathbf{s}}$ che definiscono lo stimatore (14.95) richiede quindi la soluzione del seguente problema di ottimo vincolato:

$$\begin{cases} \textit{minimizzare}: \ \sum_{i \in \mathbf{s}} \frac{(c_{i\mathbf{s}} - \frac{1}{\pi_i})^2}{\frac{q_i}{\pi_i}} \\ \textit{con i vincoli}: \ \frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}}\, x_{ik} = \mu_{x_k}; \quad k = 1, \ldots, p. \end{cases} \tag{14.109}$$

Per agevolare la soluzione del problema (14.109) è opportuno ricorrere ad una notazione vettoriale. Siano

$$t_{HTx_k} = \frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}x_{ik},\quad t_{lx_k} = \frac{1}{N}\sum_{i\in\mathbf{s}}c_{i\mathbf{s}}x_{ik};\quad k=1,\ldots,p$$

rispettivamente lo stimatore di Horvitz-Thompson e lo stimatore lineare (14.95) di μ_{x_k} $(k=1,\ldots,p)$, e siano

$$\mathbf{t}_{HTx} = \begin{bmatrix} t_{HTx_1} \\ t_{HTx_2} \\ \cdots \\ t_{HTx_p} \end{bmatrix},\quad \mathbf{t}_{lx} = \begin{bmatrix} t_{lx_1} \\ t_{lx_2} \\ \cdots \\ t_{lx_p} \end{bmatrix},\quad \boldsymbol{\mu}_x = \begin{bmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \cdots \\ \mu_{x_p} \end{bmatrix}$$

i vettori (a p componenti) rispettivamente degli stimatori di Horvitz-Thompson t_{HTx_k}, degli stimatori lineari (14.95), e delle medie μ_{x_k} $(k=1,\ldots,p)$. Indichiamo inoltre con

$$\mathbf{x}_{i.} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{ip} \end{bmatrix};\quad i=1,\ldots,N$$

il vettore (a p componenti) dei valori assunti dalle variabili ausiliarie in corrispondenza dell'unità i. Valgono le ovvie relazioni:

$$\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}\mathbf{x}_{i.} = \mathbf{t}_{HTx},\quad \frac{1}{N}\sum_{i\in\mathbf{s}}c_{i\mathbf{s}}\mathbf{x}_{i.} = \mathbf{t}_{lx}.$$

Proposizione 14.13. *La soluzione del problema di minimo vincolato* (14.109) *è costituita dai pesi:*

$$c_{i\mathbf{s}} = \frac{1}{\pi_i} - (\mathbf{t}_{HTx} - \boldsymbol{\mu}_x)'\left(\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{q_i}{\pi_i}(\mathbf{x}_{i.}\mathbf{x}'_{i.})\right)^{-1}\frac{q_i}{\pi_i}\mathbf{x}_{i.}. \tag{14.110}$$

Dimostrazione. Il problema di minimo vincolato (14.109) si risolve con il metodo dei moltiplicatori di Lagrange. La funzione Lagrangiana è pari a

$$\mathcal{L} = \sum_{i\in\mathbf{s}}\frac{(c_{i\mathbf{s}} - \frac{1}{\pi_i})^2}{\frac{q_i}{\pi_i}} - 2\boldsymbol{\lambda}'(\mathbf{t}_{lx} - \boldsymbol{\mu}_x) \tag{14.111}$$

dove $\boldsymbol{\lambda}$ è il vettore dei p moltiplicatori di Lagrange $\lambda_1,\ldots,\lambda_p$. Derivando la (14.111) rispetto ai termini $c_{i\mathbf{s}}$, e annullando tali derivate, si ha

$$\frac{\partial\mathcal{L}}{\partial c_{i\mathbf{s}}} = 2\frac{c_{i\mathbf{s}} - \frac{1}{\pi_i}}{\frac{q_i}{\pi_i}} - 2\boldsymbol{\lambda}'\mathbf{x}_{i.} = 0$$

da cui si ottiene

$$c_{i\mathbf{s}} = \frac{1}{\pi_i} + \frac{q_i}{\pi_i}\,\boldsymbol{\lambda}'\mathbf{x}_{i.}. \tag{14.112}$$

Il valore di $\boldsymbol{\lambda}$ in (14.112) si calcola utilizzando i vincoli di calibrazione. Precisamente, moltiplicando ambo i membri della (14.112) per $(1/N)\mathbf{x}_{i.}$ e sommando rispetto alle unità del campione, si ha

$$\mathbf{t}_{lx} = \mathbf{t}_{HTx} + \frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\boldsymbol{\lambda}'\mathbf{x}_{i.}\mathbf{x}_{i.}$$

da cui, tenendo conto dei p vincoli di calibrazione $\mathbf{t}_{lx} = \boldsymbol{\mu}_x$ e della relazione

$$(\boldsymbol{\lambda}'\mathbf{x}_{i.})\mathbf{x}_{i.} = (\mathbf{x}'_{i.}\boldsymbol{\lambda})\mathbf{x}_{i.} = \mathbf{x}_{i.}(\mathbf{x}'_{i.}\boldsymbol{\lambda}) = (\mathbf{x}_{i.}\mathbf{x}'_{i.})\boldsymbol{\lambda}$$

si ottiene

$$\boldsymbol{\mu}_x = \mathbf{t}_{HTx} + \left(\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\mathbf{x}_{i.}\mathbf{x}'_{i.}\right)\boldsymbol{\lambda}$$

e quindi

$$\boldsymbol{\lambda} = \left(\frac{1}{N}\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\mathbf{x}_{i.}\mathbf{x}'_{i.}\right)^{-1}(\boldsymbol{\mu}_x - \mathbf{t}_{HTx}). \tag{14.113}$$

Sostituendo la (14.113) nella (14.112) si ottiene facilmente la (14.110). □

Lo *stimatore per calibrazione* di μ_y, basato sui pesi (14.110), assume la forma:

$$t_{cal} = t_{HTy} - (\mathbf{t}_{HTx} - \boldsymbol{\mu}_x)'\left(\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,(\mathbf{x}_{i.}\mathbf{x}'_{i.})\right)^{-1}\left(\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\mathbf{x}_{i.}y_i\right). \tag{14.114}$$

Esempio 14.19 (Post-stratificazione). Supponiamo che la popolazione sia suddivisa in M strati, rispettivamente di $N_1, \ldots, N_M$ unità; al solito, indicheremo con $w_g = N_g/N$ il peso dello strato g $(= 1, \ldots, M)$. Assumiamo inoltre di conoscere i numeri $N_1, \ldots, N_M$, ma non quali unità compongono gli strati. È questa, come già visto nel Capitolo 8, la situazione tipica della post stratificazione. Sulle unità della popolazione definiamo poi le M variabili di appartenenza /non appartenenza ai diversi strati. Formalmente, poniamo

$$x_{ig} = \begin{cases} 1 \textit{ se l'unità } i \textit{ appartiene allo strato } g \\ 0 \textit{ altrimenti} \end{cases};\ i = 1, \ldots, N;\ g = 1, \ldots, M$$

e definiamo i vettori (a M componenti, di cui una eguale a 1 e le altre a 0)

$$\mathbf{x}_{i.} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \cdots \\ x_{iM} \end{bmatrix};\ i = 1, \ldots, N\,.$$

Sia inoltre $\mathbf{s}_g$ $(g = 1, \ldots, M)$ il sottocampione formato dalle unità del campione $\mathbf{s}$ appartenenti allo strato g, così che il campione "totale" $\mathbf{s}$ può essere scritto come $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M)$. Per semplicità supporremo che ognuno dei sottocampioni $\mathbf{s}_g$ contenga almeno un'unità, ossia che il campione $\mathbf{s}$ contenga almeno un'unità di ciascuno strato.

Per determinare i pesi che definiscono lo stimatore per calibrazione (14.95), iniziamo con l'osservare che i vincoli di calibrazione si scrivono come

$$\frac{1}{N} \sum_{i \in \mathbf{s}_g} c_{i\mathbf{s}} = w_g; \quad g = 1, \ldots, M$$

ovvero

$$\sum_{i \in \mathbf{s}_g} c_{i\mathbf{s}} = N_g; \quad g = 1, \ldots, M. \tag{14.115}$$

Interpretando il peso $c_{i\mathbf{s}}$ come il numero di unità della popolazione "rappresentate" dall'unità campionaria i, il termine $\sum_{i \in \mathbf{s}_g} c_{i\mathbf{s}}$ è null'altro che il numero di unità dello strato g rappresentate da quelle del campione $\mathbf{s}$. I vincoli di calibrazione (14.115) ci dicono che il numero di unità di ciascuno strato rappresentate da quelle del campione deve essere uguale al numero effettivo di unità dello strato stesso. In questo senso l'operazione di calibrazione equivale ad una *post-stratificazione*.

In secondo luogo, con ovvia notazione valgono le seguenti relazioni:

$$\boldsymbol{\mu}_x = \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_M \end{bmatrix}, \quad \mathbf{t}_{HTx} = \frac{1}{N} \begin{bmatrix} \sum_{i \in \mathbf{s}_1} \frac{1}{\pi_i} \\ \sum_{i \in \mathbf{s}_2} \frac{1}{\pi_i} \\ \cdots \\ \sum_{i \in \mathbf{s}_M} \frac{1}{\pi_i} \end{bmatrix}, \quad t_{HTy} = \frac{1}{N} \sum_{g=1}^{M} \sum_{i \in \mathbf{s}_g} \frac{1}{\pi_i} y_i.$$

Per la costruzione effettiva dello stimatore di calibrazione è necessaria qualche ulteriore osservazione. In primo luogo, se l'unità i appartiene allo strato g il prodotto $\mathbf{x}_{i.}\mathbf{x}'_{i.}$ è una matrice quadrata $M \times M$, in cui il g-mo elemento è uguale a 1 e tutti gli altri sono pari a 0. Posto $q_i = 1$ per ciascuna unità della popolazione, ne consegue che

$$\begin{aligned} \sum_{i \in \mathbf{s}} \frac{q_i}{\pi_i} (\mathbf{x}_{i.}\mathbf{x}'_{i.}) &= \sum_{g=1}^{M} \sum_{i \in \mathbf{s}_g} \frac{1}{\pi_i} (\mathbf{x}_{i.}\mathbf{x}'_{i.}) \\ &= \begin{bmatrix} 1 \Big/ \sum_{i \in \mathbf{s}_1} \frac{1}{\pi_i} & 0 & \cdots & 0 \\ 0 & 1 \Big/ \sum_{i \in \mathbf{s}_2} \frac{1}{\pi_i} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \Big/ \sum_{i \in \mathbf{s}_M} \frac{1}{\pi_i} \end{bmatrix} \end{aligned}$$

e

$$\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\mathbf{x}_{i.}y_i = \sum_{g=1}^{M}\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}\,\mathbf{x}_{i.}y_i = \begin{bmatrix} \sum_{i\in\mathbf{s}_1} \frac{1}{\pi_i}\,y_i \\ \sum_{i\in\mathbf{s}_2} \frac{1}{\pi_i}\,y_i \\ \cdots \\ \sum_{i\in\mathbf{s}_M} \frac{1}{\pi_i}\,y_i \end{bmatrix}$$

da cui si ottiene

$$\begin{aligned} t_{cal} &= t_{HTy} - (\mathbf{t}_{HTx} - \boldsymbol{\mu}_x)' \left(\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,(\mathbf{x}_{i.}\mathbf{x}_{i.}')\right)^{-1} \left(\sum_{i\in\mathbf{s}} \frac{q_i}{\pi_i}\,\mathbf{x}_{i.}y_i\right) \\ &= \frac{1}{N}\sum_{g=1}^{M}\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}\,y_i - \frac{1}{N}\sum_{g=1}^{M}\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}\,y_i + \sum_{g=1}^{M} w_g \frac{\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}\,y_i}{\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}} \\ &= \frac{1}{N}\sum_{g=1}^{M}\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}\left(\frac{N_g}{\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}}\right) y_i. \end{aligned} \tag{14.116}$$

La forma dello stimatore per calibrazione (14.116) è particolarmente significativa, e merita alcuni commenti. Essendo i pesi $c_{i\mathbf{s}}$ pari a

$$c_{i\mathbf{s}} = \frac{1}{\pi_i}\left(\frac{N_g}{\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}}\right) \textit{ per ciascuna unità campionaria dello strato } g \tag{14.117}$$

lo stimatore (14.116) "aggiusta" i propri pesi essenzialmente in modo da realizzare una post-stratificazione mediante il soddisfacimento dei vincoli (14.115). Questa interpretazione è rafforzata da una semplice considerazione. Il termine $1/\pi_i$ è il peso che il disegno di campionamento assegna all'unità i. Il termine $N_g/\sum_{i\in\mathbf{s}_g} \frac{1}{\pi_i}$ è invece un "fattore di aggiustamento", che modifica i pesi da disegno in modo da soddisfare i vincoli di calibrazione, ossia in modo da post stratificare le unità campionarie. I pesi (14.117) hanno quindi una struttura del tipo

$$\begin{array}{c}\textit{Peso assegnato dal disegno}\\ \textit{all'unità } i\end{array} \times \begin{array}{c}\textit{Fattore di aggiustamento}\\ \textit{per la post-stratificazione}.\end{array}$$ □

Esempio 14.20 (Tabelle di contingenza con vincoli sulle marginali). Si supponga che sulle unità della popolazione siano definiti due caratteri discreti A, B, aventi modalità categoriali rispettivamente $A_1, \ldots, A_H$, $B_1, \ldots, B_K$. Per ciascuna unità della popolazione definiamo poi le $H + K$ variabili

$$x_{ih}^{A} = \begin{cases} 1 & \textit{l'unità } i \textit{ possiede modalità } A_h \\ 0 & \textit{altrimenti} \end{cases};$$

$$x_{ik}^{B} = \begin{cases} 1 & \textit{l'unità } i \textit{ possiede modalità } B_k \\ 0 & \textit{altrimenti} \end{cases}$$

$(i = 1, \ldots, N;\ h = 1, \ldots, H;\ k = 1, \ldots, K)$.

Supponiamo inoltre di voler stimare la proporzione di unità della popolazione che possiedono una qualsiasi coppia (A_h, B_k) di modalità dei due caratteri $(h = 1, \ldots, H;\ k = 1, \ldots, K)$. Se si definiscono le HK variabili di interesse

$$y_{ihk} = x_{ih}^A\, x_{ik}^B = \begin{cases} 1 \ l'unità\ i\ possiede\ modalità\ A_h\ e\ B_k \\ 0\ altrimenti \end{cases}$$

$(i = 1, \ldots, N;\ h = 1, \ldots, H;\ k = 1, \ldots, K)$ le proporzioni da stimare sono

$$P_{hk} = \frac{1}{N} \sum_{i=1}^{N} y_{ihk} = \frac{N_{hk}}{N};\ \ h = 1, \ldots, H;\ k = 1, \ldots, K \qquad (14.118)$$

essendo N_{hk} il numero di unità che possiedono le modalità A_h e B_k.

Lo stimatore di Horvitz-Thompson di P_{hk} è uguale a

$$\widehat{P}_{hk}^{HT} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i}\, y_{ihk};\ \ h = 1, \ldots, H;\ k = 1, \ldots, K. \qquad (14.119)$$

Supponiamo ora che siano note le proporzioni

$$P_{h.} = \sum_{k=1}^{K} P_{hk};\ \ h = 1, \ldots, H$$

di unità che possiedono le modalità $A_1, \ldots, A_H$, e le proporzioni

$$P_{.k} = \sum_{h=1}^{H} P_{hk};\ \ k = 1, \ldots, K$$

di unità che possiedono le modalità $B_1, \ldots, B_K$.

Se indichiamo con $\widehat{P}_{hk}^{cal}$ lo stimatore per calibrazione di P_{hk}, dovranno essere soddisfatti i vincoli

$$\sum_{h=1}^{H} \widehat{P}_{hk}^{cal} = P_{h.},\ \ h = 1, \ldots, H \qquad (14.120)$$

$$\sum_{k=1}^{K} \widehat{P}_{hk}^{cal} = P_{.k},\ \ k = 1, \ldots, K. \qquad (14.121)$$

Gli stimatori $\widehat{P}_{hk}^{cal}$, per $h = 1, \ldots, H$, $k = 1, \ldots, K$ possono essere costruiti lungo le linee indicate in precedenza. Nel caso in cui le q_i siano tutte eguali a 1 e le stime $\widehat{P}_{hk}^{HT}$ siano tutte positive, è però possibile percorrere una strada alternativa, molto interessante, basata sull'algoritmo di *Iterative Proportional Fitting* (IFP) di seguito brevemente descritto.

– *Passo* 0. Inizializzazione. Porre $t = 0$ e $\widehat{P}_{hk}^{(0)} = \widehat{P}_{hk}^{HT}$, per $h = 1, \ldots, H$, $k = 1, \ldots, K$. Fissare un ''livello di soglia di arresto'' $\delta > 0$.

- *Passo* 1. Aggiustamento delle marginali di riga. Porre

$$\widehat{P}_{hk}^{(t+1)} = \widehat{P}_{hk}^{(t)} \frac{P_{h.}}{\sum_{k=1}^{K} \widehat{P}_{hk}^{(t)}}$$

per $h = 1, \ldots, H$, $k = 1, \ldots, K$. Incrementare t di 1 e andare al Passo 2.
- *Passo* 2. Aggiustamento delle marginali di colonna. Porre

$$\widehat{P}_{hk}^{(t+1)} = \widehat{P}_{hk}^{(t)} \frac{P_{.k}}{\sum_{h=1}^{H} \widehat{P}_{hk}^{(t)}}$$

per $h = 1, \ldots, H$, $k = 1, \ldots, K$. Incrementare t di 1 e andare al Passo 3.
- *Passo* 3. Verifica della condizione di arresto. Se

$$|\widehat{P}_{hk}^{(t)} - \widehat{P}_{hk}^{(t-2)}| < \delta$$

per ciascun $h = 1, \ldots, H$, $k = 1, \ldots, K$, andare al Passo 4. Altrimenti, andare al Passo 1.
- *Passo* 4. Arresto. Porre $\widehat{P}_{hk}^{cal} = \widehat{P}_{hk}^{(t)}$ per ciascun $h = 1, \ldots, H$, $k = 1, \ldots, K$.

□

Esempio 14.21 (Stimatore per regressione). Supponiamo che il disegno campionario sia semplice senza ripetizione (così che $\pi_i = n/N$), e sia $\mathcal{X}$ una variabile ausiliaria con media μ_x nota, ed i cui valori x_i siano (almeno) osservati sulle unità campionarie. Lo stimatore di Horvitz-Thompson di μ_y è uguale, ovviamente, alla media campionaria $\overline{y}_{\mathbf{s}}$. Per adottare il simbolismo usato in precedenza, definiamo le due variabili $\mathcal{X}_1$, $\mathcal{X}_2$ come

$$x_{i1} = 1,\ x_{i2} = x_i;\ \ i = 1, \ldots, N.$$

Chiaramente, le medie di queste due variabili (sull'intera popolazione) sono $\mu_{x_1} = 1$ e $\mu_{x_2} = \mu_x$.

Se $c_{i\mathbf{s}}$ sono i pesi dello stimatore per calibrazione, devono essere soddisfatti i vincoli:

$$\frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}} = 1,\ \ \frac{1}{N} \sum_{i \in \mathbf{s}} c_{i\mathbf{s}} x_i = \mu_x.$$

Per quanto concerne la costruzione dello stimatore per calibrazione, osserviamo anzitutto che i vettori (a due componenti) $\mathbf{x}_{i.}$ sono del tipo

$$\mathbf{x}_{i.} = \begin{bmatrix} 1 \\ x_i \end{bmatrix};\ \ i = 1, \ldots, N$$

così che si ha

$$\left(\frac{1}{n}\sum_{i\in\mathbf{s}}\mathbf{x}_{i.}\mathbf{x}'_{i.}\right)^{-1}\left(\frac{1}{n}\sum_{i\in\mathbf{s}}\mathbf{x}_{i.}\,y_i\right)=\begin{bmatrix}\overline{y}_{\mathbf{s}}-\widehat{b}_{y/x}\overline{x}_{\mathbf{s}}\\ \widehat{b}_{y/x}\end{bmatrix}$$

essendo

$$\widehat{b}_{y/x}=\frac{\sum_{i\in\mathbf{s}}x_i\,y_i/n-\overline{x}_{\mathbf{s}}\overline{y}_{\mathbf{s}}}{\sum_{i\in\mathbf{s}}x_i^2/n-\overline{x}_{\mathbf{s}}^2}$$

il coefficiente di regressione campionario di $\mathcal{Y}$ rispetto a $\mathcal{X}$.

Lo stimatore t_{cal} assume la forma

$$t_{cal}=\overline{y}_{\mathbf{s}}-\widehat{b}_{y/x}(\overline{x}_{\mathbf{s}}-\mu_x)$$

e si riduce quindi al classico stimatore per regressione introdotto nel Capitolo 5. □

Esercizi

14.1. Provare che lo stimatore per regressione si può scrivere nella forma (14.11).

14.2. Provare che lo stimatore per quoziente è corretto se si usa un disegno di Midzuno-Lahiri con $p_i = x_i/(N\mu_x)$.

14.3. Mostrare che se il disegno campionario è di tipo sistematico, non esiste nessuno stimatore corretto di $V(t_{HT})$ della forma:

$$\widehat{V}=\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}c_{ij}\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}$$

essendo c_{ij} opportuni coefficienti numerici.

14.4. Provare che se $\pi_{ij} > 0$ per tutte le coppie i, j di unità distinte, l'unico stimatore corretto di $V(t_{HT})$ della forma:

$$\widehat{V}=\frac{1}{N^2}\sum_{i\in\mathbf{s}}\sum_{j\in\mathbf{s}}c_{ij}\,\frac{y_i}{\pi_i}\,\frac{y_j}{\pi_j}$$

si ottiene per $c_{ij} = \Delta_{ij}/\pi_{ij}$.

14.5. Data una popolazione finita di N unità, si considerino due caratteri $\mathcal{X}$, $\mathcal{Y}$, che assumono rispettivamente le modalità $x_1, \ldots, x_N$ e $y_1, \ldots, y_N$. Dette μ_x, μ_y le medie dei due caratteri, si considerino i loro stimatori di Horvitz-Thompson

$$t_{HTx}=\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}x_i,\quad t_{HTy}=\frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}y_i.$$

Provare che la covarianza tra t_{HTx} e t_{HTy} è uguale a

$$C(t_{HTx}, t_{HTy}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{x_i}{\pi_i} \frac{y_j}{\pi_j} \Delta_{ij}$$

con $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$.

14.6. Provare la relazione (14.52).

14.7. Provare la relazione (14.54).

14.8. Provare che lo stimatore (14.51) non è migliorabile con la tecnica di contrazione.

14.9. Provare che lo stimatore (14.53) non è migliorabile con la tecnica di contrazione.

14.10. (Stimatore di Horvitz-Thompson per disegni *unicluster*) Un disegno campionario $(\mathcal{S}, p(\cdot))$ è detto *unicluster* se i campioni in $\mathcal{S}$ sono due a due disgiunti, ossia se $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K\}$ con $\mathbf{s}_j \cap \mathbf{s}_l = \emptyset$ per ogni $j \neq l$, $j, l = 1, \ldots, K$. Il più importante tipo di disegno *unicluster* è il disegno sistematico. Verificare che per questo tipo di disegno lo stimatore di Horvitz-Thompson di μ_y non è migliorabile con la tecnica di contrazione.

14.11. (Stimatore di Horvitz-Thompson e disegno di Bernoulli) Dato un (arbitrario) numero $0 < p < 1$, siano $B_1, \ldots, B_N$ N variabili aleatorie indipendenti, identicamente distribuite, e tali che $Pr(B_i) = 1 = p$, $Pr(B_i = 0) = 1 - p$, per ciascun $i = 1, \ldots, N$. Si consideri il disegno campionario definito da $\delta(i; \mathbf{s}) = B_i$ per ciascun $i = 1, \ldots, N$ (l'unità i è inclusa nel campione $\mathbf{s}$ se e solo se $B_i = 1$).

a. Verificare che $\pi_i = p$ per ciascun $i = 1, \ldots, N$, e che $p_{ij} = p^2$ per ciascun $j \neq i$.

b. Verificare che lo stimatore di Horvitz-Thompson di μ_y,

$$t_{HT} = \frac{1}{Np} \sum_{i=1}^{N} y_i \, \delta(i; \mathbf{s}),$$

ha varianza

$$V(t_{HT}) = \frac{1}{N^2} \frac{1-p}{p} \sum_{i=1}^{N} y_i^2.$$

c. Verificare che il rapporto $E[t_{HT}^2]/\mu_y^2 = 1 + V(t_{HT})/\mu_y^2$, qualunque sia μ_y fissato, raggiunge il suo valore minimo per $y_1 = y_2 = \cdots = y_N$ $(= \mu_y)$.

d. Verifiare che il valore ottimo della costante di contrazione è $c^* = \frac{Np}{(N-1)p+1}$.

14.12. Siano $\mathcal{X}$, $\mathcal{Y}$ due caratteri di interesse, di cui si vogliono stimare le medie μ_x, μ_y, e siano

$$t_{HTx} = \frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}x_i, \quad t_{HTy} = \frac{1}{N}\sum_{i\in\mathbf{s}}\frac{1}{\pi_i}y_i$$

i corrispondenti stimatori di Horvitz-Thompson. Verificare che la loro covarianza è eguale a:

$$C(t_{HTx},\, t_{HTy}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{x_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij} - \pi_i\pi_j).$$

14.13. Dimostrare che nel caso di disegno ssr di ampiezza n, e se $y_i^0 = cx_i$ per ciascun $i = 1, \ldots, N$, lo stimatore alle differenze generalizzate ha (con ovvia simbologia) varianza pari a

$$V(t_{gd}) = \left(\frac{1}{n} - \frac{1}{N}\right)(S_y^2 + S_x^2 - 2S_{xy}). \tag{14.122}$$

14.14. Provare le relazioni (14.81), (14.82).

14.15. Verificare la relazione (14.94).

15

Costruzione di disegni campionari con preassegnate caratteristiche

15.1 Aspetti introduttivi. Qualità "desiderabili" di disegni campionari

L'obiettivo di questo capitolo è quello di esporre alcuni disegni campionari che possiedono "buone" proprietà, che li rendono particolarmente importanti per le loro potenzialità applicative. Una particolare attenzione sarà dedicata agli aspetti algoritmici, in particolare per quel che riguarda la definizione di schemi computazionalmente efficienti per selezionare campioni di unità sulla base di un dato disegno di campionamento (*implementazione* del disegno campionario). Si tratta di un punto molto importante, in quanto l'uso effettivo di un disegno campionario è legato in modo indissolubile alla possibilità di implementarlo in modo numericamente efficiente, ossia alla possibilità di selezionare in modo numericamente efficiente un campione.

La scelta di un disegno campionario non è un problema a sé stante, ma va sempre visto in coppia con la scelta dello stimatore del parametro di interesse (o degli stimatori dei parametri di interesse). Ha quindi più senso parlare di scelta di una *strategia di campionamento*, ossia di una coppia (disegno, stimatore). Se l'obiettivo è quello di stimare un certo parametro, il criterio-guida dovrebbe essere quello di scegliere una strategia di campionamento altamente efficiente, ossia caratterizzata da un "piccolo" errore quadratico medio di stima.

In questo capitolo largo spazio sarà dedicato a disegni campionari da utilizzare in coppia con lo stimatore di Horvitz-Thompson della media della popolazione. Come già si è avuto modo di chiarire, per garantire buone proprietà di efficienza allo stimatore di Horvitz-Thompson è necessario che il disegno di campionamento abbia ampiezza effettiva costante, e che le probabilità di inclusione del primo ordine siano, in via approssimata, proporzionali alle modalità y_i. Essendo queste ultime incognite, di fatto si useranno probabilità di inclusione del primo ordine proporzionali alle modalità x_i di una variabile ausiliaria $\mathcal{X}$, come descritto nella Sezione 14.2.5. Quanto più tra le y_i e

Conti P. L., Marella D.: Campionamento da popolazioni finite. Il disegno campionario.
DOI 10.1007/978-88-470-2577-6_15,

le x_i sussiste una relazione "vicina" alla proporzionalità, tanto migliore sarà questa scelta. Indicheremo con π_{0i} tali probabilità di inclusione "desiderate". Chiaramente, detta n la numerosità campionaria, ciò significa porre $\pi_{0i} = 1$ per tutte le unità i tali che $nx_i / \sum_{j=1}^{N} x_j \geqslant 1$, e ricalcolare le probabilità di inclusione delle restanti unità come

$$\pi_{0i} = (n - n_A) \frac{x_i}{\sum_{j=1 j \notin A}^{N} x_j}; \quad i = 1, \ldots, N; \; i \notin A$$

essendo A l'insieme delle unità della popolazione tali che $nx_i / \sum_{j=1}^{N} x_j \geqslant 1$, e n_A il numero di tali unità. In questo modo, alla fine si avranno probabilità di inclusione "desiderate" π_{0i} tutte minori o uguali a 1 e la cui somma è pari a n.

Per garantire buone proprietà di efficienza allo stimatore di Horvitz-Thompson della media della popolazione, il disegno campionario dovrebbe avere (almeno) le seguenti due caratteristiche.

$C1$. Ampiezza effettiva costante n.
$C2$. Probabilità di inclusione del primo ordine pari a quelle desiderate π_{01}, $\ldots$, π_{0N}.

Naturalmente, vi sono anche altre proprietà che un disegno dovrebbe possedere. Ad esempio, le probabilità di inclusione del secondo ordine dovrebbero essere calcolabili in modo efficiente almeno in via approssimata, in modo da poter stimare la varianza dello stimatore di Horvitz-Thompson; inoltre, tale stimatore dovrebbe essere non negativo. Requisiti aggiuntivi rispetto a $C1$ e $C2$ sono elencati di seguito.

$C3$. Varianza dello stimatore di Horvitz-Thompson "piccola".
$C4$. Probabilità di inclusione del secondo ordine calcolabili in modo numericamente efficiente, almeno in via approssimata. Questo requisito è importante per stimare la varianza dello stimatore di Horvitz-Thompson. Da questo punto di vista, le probabilità di inclusione del secondo ordine dovrebbero soddisfare la disuguaglianza $\pi_{ij} \leqslant \pi_i \pi_j$ per tutte le coppie di unità distinte i, j, in modo che lo stimatore di Yates-Grundy di $V(t_{HT})$ sia non negativo.
$C5$. Il disegno campionario dovrebbe possedere un'entropia "sufficientemente grande", in modo da avere una certa robustezza di stima nel caso di popolazione "non troppo vicina" alla situazione ideale in cui le y_i sono proporzionali alle x_i.

Non tutti i requisiti $C1$–$C5$ possono essere simultaneamente soddisfatti in modo esatto. Ad esempio, per quanto riguarda $C3$ è possibile verificare (Esercizio 15.1) che tra tutti i disegni ad ampiezza effettiva costante e prefissate probabilità di inclusione del primo ordine, non ne esiste uno che rende minima la varianza dello stimatore di Horvitz-Thompson $V(t_{HT})$. Quindi, il requisito $C3$ in assoluto non è realmente perseguibile nel senso di determinare le pro-

babilità di inclusione del secondo ordine che rendono minima la varianza dello stimatore di Horvitz-Thompson.

La proprietà $C5$ (entropia "grande") è importante non solo di per sé, ma anche perché permette di costruire approssimazioni "ragionevoli" delle probabilità di inclusione del secondo ordine, così come della varianza $V(t_{HT})$.

Di schemi campionari per la selezione di campioni con preassegnate caratteristiche ne esistono moltissimi in letteratura, e non è nostro obiettivo fornire neanche un loro sommario elenco. Il lettore interessato può consultare i volumi di Brewer e Hanif (1983) e di Chaudhuri e Vos (1988). Ci si limiterà essenzialmente a descrivere alcuni dei più importanti schemi di campionamento, cercando di mettere in evidenza gli aspetti algoritmici legati al loro uso effettivo. Un'eccellente monografia dedicata a questi aspetti è il bel volume di Tillé (2006).

15.2 Disegni campionari di Poisson e di Bernoulli

15.2.1 Il disegno campionario di Poisson

Un disegno di campionamento con probabilità variabili e avente struttura molto semplice è il disegno di Poisson. Siano $p_1, \ldots, p_N$ N numeri tali che

$$0 < p_i \leqslant 1 \;\; \textit{per ciascuna unità } i = 1, \ldots N;$$
$$p_1 + \cdots + p_N = 1.$$

Nel *disegno campionario di Poisson* si assume che le variabili indicatrici $\delta(i; \mathbf{s})$ sono *indipendenti* e tali che

$$P(\delta(i, \mathbf{s}) = 1) = p_i \;\; \textit{per ciascuna unità } i = 1, \ldots, N. \tag{15.1}$$

Per quanto riguarda lo spazio dei campioni, poiché ciascun indicatore $\delta(i; \mathbf{s})$ può assumere in modo indipendente i valori 0, 1, lo spazio campionario è costituito dai 2^N sottoinsiemi di $\{1, \ldots, N\}$. Inoltre, un qualunque campione $\mathbf{s}$ ha probabilità

$$\begin{aligned} p(\mathbf{s}) &= Pr(\delta(i; \mathbf{s}) = 1 \, per \; i \in \mathbf{s};\; \delta(i; \mathbf{s}) = 0 \, per \; i \notin \mathbf{s}) \\ &= \left\{ \prod_{i \in \mathbf{s}} p_i \right\} \left\{ \prod_{i \notin \mathbf{s}} (1 - p_i) \right\} \\ &= \prod_{i=1}^{N} p_i^{\delta(i; \mathbf{s})} (1 - p_i)^{1 - \delta(i; \mathbf{s})}. \end{aligned} \tag{15.2}$$

La (15.2) può anche scriversi nella forma

$$p(\mathbf{s}) = C_{po} \prod_{i=1}^{N} \omega_i^{\delta(i; \mathbf{s})} \tag{15.3}$$

dove si è posto

$$\omega_i = \frac{p_i}{1-p_i} \quad \textit{per ciascuna unità } i = 1, \ldots, N \tag{15.4}$$

$$C_{po} = \prod_{i=1}^{N}(1-p_i). \tag{15.5}$$

Il calcolo delle probabilità di inclusione del primo e del secondo ordine è semplicissimo. Dalla (15.1) si ha in primo luogo:

$$\pi_i = p_i \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{15.6}$$

Dall'indipendenza delle $\delta(i; \mathbf{s})$ si ha poi:

$$\begin{aligned} \pi_{ij} &= Pr(\delta(i; \mathbf{s}) = 1\ \delta(j; \mathbf{s}) = 1) \\ &= Pr(\delta(i; \mathbf{s}) = 1)\, Pr(\delta(j; \mathbf{s}) = 1) \\ &= p_i\, p_j \end{aligned} \tag{15.7}$$

per ogni coppia i, j di unità distinte.

Le relazioni appena trovate permettono di risolvere con molta facilità un importante problema: "Quali valori devono assumere $p_1, \ldots, p_N$ in modo che le probabilità di inclusione del primo ordine siano esattamente uguali a $\pi_{01}, \ldots, \pi_{0N}$?" Dalla (15.6) risulta immediato che si deve porre:

$$p_1 = \pi_{01},\ p_2 = \pi_{02}, \ldots, p_N = \pi_{0N}. \tag{15.8}$$

In altri termini, per soddisfare il requisito $C2$ è sufficiente scegliere i numeri $p_1, \ldots, p_N$ in base alla (15.8). Con questa scelta, le probabilità di inclusione del secondo ordine sono : $\pi_{ij} = \pi_{0i}\, \pi_{0j}$ per tutte le coppie i, j di unità distinte.

Molto facile è anche la generazione di un campione in base al disegno di Poisson, che si può realizzare mediante un facile schema sequenziale. È sufficiente *generare N variabili aleatorie $U_1, \ldots, U_N$ indipendenti e tutte con distribuzione uniforme in* $[0, 1]$. *Se $U_i \leqslant p_i$ l'unità i entra a far parte del campione (e si pone $\delta(i; \mathbf{s}) = 1$); se $U_i > p_i$ l'unità i non entra a far parte del campione (e si pone $\delta(i; \mathbf{s}) = 0$).*

Un'ultima proprietà positiva, e importante, del disegno di Poisson riguarda la sua entropia. È infatti possibile dimostrare (Esercizio 15.2) che *tra tutti i disegni campionari non ordinati, senza ripetizioni, e con prefissate probabilità di inclusione del primo ordine $\pi_{01}, \ldots, \pi_{0N}$, il disegno di Poisson è quello di entropia massima.* Il requisito $C5$ è quindi soddisfatto.

Malgrado tutte le proprietà che possiede, il disegno di Poisson ha un difetto che lo rende praticamente inutilizzabile assieme allo stimatore di Horvitz-Thompson: dà luogo a campioni di dimensione variabile, da 0 a N. In particolare, $n(\mathbf{s}) = 0$ corrisponde al caso $\delta(i; \mathbf{s}) = 0$ per tutte le unità della popolazione (ed ha luogo con probabilità $\prod(1-\pi_i)^N$), mentre $n(\mathbf{s}) = N$ corrisponde al caso $\delta(i; \mathbf{s}) = 1$ per tutte le unità della popolazione (ed ha luogo

con probabilità ($\prod \pi_i^N$). Formalmente, la numerosità campionaria è pari a

$$n(\mathbf{s}) = \sum_{i=1}^{N} \delta(i;\, \mathbf{s}). \tag{15.9}$$

Sulla base dei risultati introdotti nel Capitolo 12 si può facilmente dimostrare che il valore atteso e la varianza della numerosità campionaria $n(\mathbf{s})$ sono nel nostro caso pari a

$$E[n(\mathbf{s})] = \sum_{i=1}^{N} p_i; \tag{15.10}$$

$$V(n(\mathbf{s})) = \sum_{i=1}^{N} p_i(1 - p_i). \tag{15.11}$$

La distribuzione di probabilità di $n(\mathbf{s})$ è quella della somma di N variabili indipendenti di Bernoulli, ma non aventi la stessa distribuzione di probabilità (a meno che non sia $p_1 = \cdots = p_N$). Ad ogni modo, detta $\mathcal{C}_{N,n}$ la famiglia di tutti i sottoinsiemi di n unità della popolazione (combinazioni senza ripetizioni di classe n), è facile vedere che

$$Pr(n(\mathbf{s}) = n) = \sum_{\mathbf{s} \in \mathcal{C}_{N,n}} p(\mathbf{s}) = C_{po} \sum_{\mathbf{s} \in \mathcal{C}_{N,n}} \prod_{i=1}^{N} \omega_i^{\delta(i;\, \mathbf{s})} \tag{15.12}$$

con i numeri ω_i definiti in (15.4). Purtroppo, a meno di casi speciali, quest'espressione non si può semplificare.

Per quanto riguarda il comportamento dello stimatore di Horvitz-Thompson t_{HT} della media della popolazione, la sua varianza assume una forma particolarmente semplice in forza della (15.7). È infatti facile verificare (Esercizio 15.3) che se il disegno campionario è di Poisson, con $p_i = \pi_{0i}$, lo stimatore di Horvitz-Thompson della media della popolazione (che ovviamente è corretto) ha varianza

$$V(t_{HT}) = \frac{1}{N^2} \sum_{i \in U} \frac{1 - \pi_{0i}}{\pi_{0i}}\, y_i^2. \tag{15.13}$$

Dall'espressione (15.13) si desume facilmente che in questo caso lo stimatore di Horvitz-Thompson è migliorabile con la tecnica di contrazione (per dettagli si veda l'Esercizio 15.4). Purtroppo, quindi, il disegno di Poisson è del tutto fuori luogo nell'ambito di strategie di campionamento che usano lo stimatore di Horvitz-Thompson. Affinché questo stimatore possa, almeno in via potenziale, fornire buoni risultati, è necessario limitarsi solo a *disegni campionari ad ampiezza effettiva costante*.

15.2.2 Il disegno campionario di Bernoulli

Il disegno campionario di Bernoulli è un caso particolare di quello di Poisson, in cui i numeri π_{01} sono tutti uguali. Formalmente, si pone $p_1 = \cdots = p_N = p$, con $0 < p < 1$. Si tratta anche in questo caso di un disegno non ordinato, senza ripetizioni, e ad ampiezza variabile. Anche qui, lo stimatore di Horvitz-Thompson di μ_y è migliorabile con la tecnica della contrazione (Esercizio 14.11).

Alcune proprietà del disegno di Bernoulli sono interessanti, e meritano di essere messe in rilievo. In primo luogo, gli indicatori $\delta(1; \mathbf{s}), \ldots, \delta(N; \mathbf{s})$, oltre ad essere indipendenti hanno anche la stessa distribuzione di probabilità (di Bernoulli di parametro p). Ne consegue che la numerosità campionaria $n(\mathbf{s}) = \sum_{i=1}^{N} \delta(i; \mathbf{s})$ ha distribuzione binomiale di parametri (N, p). In simboli

$$Pr(n(\mathbf{s}) = n) = \binom{N}{n} p^n (1-p)^{N-n}; \quad n = 0, 1, \ldots, N. \tag{15.14}$$

È anche interessante studiare il *disegno di Bernoulli condizionato* alla numerosità campionaria. Formalmente, esso è definito considerando solo i campioni di numerosità n, e calcolando la corrispondente probabilità condizionata $p(\mathbf{s}\,|n(\mathbf{s}) = n)$. Se ci si restringe ai soli campioni s di n unità, lo spazio dei campioni si riduce ovviamente a $\mathcal{C}_{N,n}$, la famiglia di tutte le combinazioni senza ripetizione di n unità della popolazione (o equivalentemente, la famiglia di tutti i sottoinsiemi di n unità). In base al disegno di Bernoulli, un campione $\mathbf{s}$ di n unità ha probabilità

$$p(\mathbf{s}) = \left\{ \prod_{i=1}^{N} p^{\delta(i;\,\mathbf{s})} \right\} \left\{ \prod_{i=1}^{N} (1-p)^{1-\delta(i;\,\mathbf{s})} \right\} = p^n (1-p)^{N-n}.$$

La probabilità del campione $\mathbf{s}$ condizionata al numero n di unità che lo compongono è quindi

$$\begin{aligned} p(\mathbf{s}\,|n(\mathbf{s}) = n) &= \frac{p(\mathbf{s})}{Pr(n(\mathbf{s}) = n)} \\ &= \frac{p^n (1-p)^{N-n}}{\binom{N}{n} p^n (1-p)^{N-n}} \\ &= \frac{1}{\binom{N}{n}} \quad \textit{per ciascun campione } \mathbf{s} \in \mathcal{C}_{N,n}\,. \end{aligned} \tag{15.15}$$

Dalla (15.15) si conclude pertanto che *il disegno di Bernoulli condizionato al numero di unità campionarie si riduce al disegno semplice senza ripetizione.*

15.3 Il disegno campionario di Sampford

15.3.1 Aspetti introduttivi e di base

Il disegno campionario di Sampford (1967) presenta notevole importanza nell'ambito del campionamento, a causa delle proprietà che possiede. L'obiettivo, come già rimarcato, è quello di costruire un disegno che possieda almeno le proprietà $C1$, $C2$ viste in precedenza.

Siano dati N numeri $p_1, \ldots, p_N$, tali che

$$0 < p_i \leqslant 1 \quad \textit{per ciascuna unità } i = 1, \ldots N;$$
$$p_1 + \cdots + p_N = n.$$

Il disegno campionario di Sampford è definito da: (a) uno spazio dei campioni (di unità) eguale alla famiglia di tutti i sottoinsiemi di n unità della popolazione: $\mathcal{S} = \mathcal{C}_{N,n}$; (b) probabilità dei campioni pari a:

$$p(\mathbf{s}) = A_s \left\{ \prod_{i=1}^{N} p_i^{\delta(i;\mathbf{s})} (1 - p_i)^{1-\delta(i;\mathbf{s})} \right\} \left(\sum_{i=1}^{N} (1 - p_i) \delta(i;\, \mathbf{s}) \right) \tag{15.16}$$

essendo A_s un'opportuna costante, tale da soddisfare la condizione che la somma (rispetto a $\mathbf{s}$ in $\mathcal{C}_{N,n}$) delle (15.16) sia pari a 1.

La (15.16) si può anche scrivere in una forma alternativa. Se, similmente al disegno di Poisson, poniamo

$$\omega_i = \frac{p_i}{1 - p_i} \quad \textit{per ciascuna unità } i = 1, \ldots, N \tag{15.17}$$

la (15.16)

$$\begin{aligned} p(\mathbf{s}) &= C_s \left\{ \prod_{i=1}^{N} \omega_i^{\delta(i;\mathbf{s})} \right\} \left(\sum_{i=1}^{N} (1 - p_i) \delta(i;\, \mathbf{s}) \right) \\ &= C_s \left\{ \prod_{i=1}^{N} \omega_i^{\delta(i;\mathbf{s})} \right\} \left(n - \sum_{i=1}^{N} p_i \delta(i;\, \mathbf{s}) \right) \end{aligned} \tag{15.18}$$

dove C_s è un'opportuna costante tale che la somma delle (15.18) sia eguale a 1.

Il calcolo esplicito della costante C_s è tutt'altro che semplice, in quanto richiede l'uso di relazioni combinatorie non banali. Sia $\mathcal{C}_{N,m}$ la classe di tutte le combinazioni senza ripetizioni di m $(= 0, 1, \ldots, N)$ unità della popolazione I_N, e siano $D_0 = 1$ e

$$D_m = \sum_{\mathbf{c} \in \mathcal{C}_{N,m}} \left(\prod_{j=1}^{N} \omega_j^{\delta(j;\mathbf{c})} \right) \tag{15.19}$$

essendo, come al solito

$$\delta(j;\,\mathbf{c}) = \begin{cases} 1 \; se \; j \in \mathbf{c} \\ 0 \; se \; j \notin \mathbf{c} \end{cases}.$$

È possibile provare (Esercizio 15.5) che vale la relazione

$$C_s = \frac{1}{\sum_{t=1}^{n} t\, D_{n-t}}. \tag{15.20}$$

Usando poi le quantità definite nell'Esercizio 15.5, è facile provare (Esercizio 15.6) che il disegno di Sampford ha probabilità di inclusione del primo ordine eguali a $p_1, \ldots, p_N$. In simboli:

$$\pi_i = p_i \;\; per \; ciascuna \; unità \; i = 1, \ldots, N. \tag{15.21}$$

Questa relazione permette di risolvere in maniera facile il (solito) problema: "Quali valori devono assumere $p_1, \ldots, p_N$ in modo che le probabilità di inclusione del primo ordine siano esattamente uguali a $\pi_{01}, \ldots, \pi_{0N}$?" La (15.21) mostra che per avere le probabilità di inclusione "desiderate" $\pi_{01}, \ldots, \pi_{0N}$ è sufficiente porre

$$p_i = \pi_{0i} \;\; per \; ciascuna \; unità \; i = 1, \ldots, N. \tag{15.22}$$

Il calcolo delle probabilità di inclusione del secondo ordine è piuttosto impegnativo, anche se non troppo difficile. Date due unità distinte i, j, sia $\mathcal{C}_{N-2,m}(\overline{i}, \overline{j})$ la classe di tutte le combinazioni senza ripetizioni di m unità della popolazione I_N privata di i e j ($I_N \setminus \{i, j\}$). Poniamo inoltre

$$D_m(\overline{i}, \overline{j}) = \sum_{\mathbf{c} \in \mathcal{C}_{N-2,m}(\overline{i}, \overline{j})} \left(\prod_{k=1}^{N} \omega_k^{\delta(k;\mathbf{c})} \right). \tag{15.23}$$

Si può verificare (Esercizio 15.7) che le probabilità di inclusione del secondo ordine per il disegno di Sampford sono eguali a

$$\pi_{ij} = C_s\, \omega_i\, \omega_j \sum_{t=2}^{n} (t - p_i - p_j) D_{n-t}(\overline{i}, \overline{j}). \tag{15.24}$$

Le probabilità di inclusione del secondo ordine dipendono dai termini $D_m(\overline{i}, \overline{j})$ in (15.24). Per calcolarli, non c'è bisogno di enumerare le combinazioni di $\mathcal{C}_{N-2,m}(\overline{i}, \overline{j})$. Infatti, è facile provare (Esercizio 15.8) che vale la relazione:

$$D_m = \frac{1}{m} \sum_{r=1}^{m} (-1)^{r-1} \left(\sum_{i=1}^{N} \omega_i^r \right) D_{m-r} \tag{15.25}$$

con $D_0 = 1$. Usando anche la relazione (semplice conseguenza della (b) dell'Esercizio 15.7)

$$D_m(\overline{i}, \overline{j}) = D_m - (\omega_i + \omega_j) D_{m-1}(\overline{i}, \overline{j}) - \omega_i \omega_j D_{m-2}(\overline{i}, \overline{j}) \qquad (15.26)$$

con $D_0(\overline{i}, \overline{j}) = 1$, è facile calcolare le (15.24) senza enumerare combinazioni, in modo quindi abbastanza efficiente sul piano computazionale.

Esempio 15.1. Nel caso di numerosità campionaria $n = 2$ si ottengono in modo facile risultati espliciti. Ogni campione è un sottoinsieme di due unità della popolazione, così che in totale vi sono $N(N-1)/2$ campioni. Se indichiamo con $\pi_{01}, \ldots, \pi_{0N}$ le probabilità di inclusione del primo ordine desiderate, il generico campione $\mathbf{s} = \{i, j\}$ ha probabilità

$$\begin{aligned} p(\mathbf{s}) &= C_s \left\{ (1 - \pi_{0i}) \frac{\pi_{0i}\, \pi_{0j}}{(1 - \pi_{0i})(1 - \pi_{0j})} + (1 - \pi_{0j}) \frac{\pi_{0j}\, \pi_{0i}}{(1 - \pi_{0j})(1 - \pi_{0i})} \right\} \\ &= C_s\, \pi_{0i}\, \pi_{0j} \left\{ \frac{1}{1 - \pi_{0i}} + \frac{1}{1 - \pi_{0j}} \right\}. \end{aligned}$$

Il calcolo della costante C_s può essere effettuato direttamente, senza usare le formule sviluppate in precedenza. Si ha:

$$\begin{aligned} \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) &= C_s \sum_{i=1}^{N} \sum_{j>i} \pi_{0i}\, \pi_{0j} \left\{ \frac{1}{1 - \pi_{0i}} + \frac{1}{1 - \pi_{0j}} \right\} \\ &= \frac{C_s}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{0i}\, \pi_{0j} \left\{ \frac{1}{1 - \pi_{0i}} + \frac{1}{1 - \pi_{0j}} \right\} \\ &= C_s \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\pi_{0i}}{1 - \pi_{0i}}\, \pi_{0j} \\ &= C_s \sum_{i=1}^{N} (2 - \pi_{0i}) \frac{\pi_{0i}}{1 - \pi_{0i}} \end{aligned}$$

e tale quantità è pari a 1 solo se

$$C_s = \frac{1}{\sum_{i=1}^{N} (2 - \pi_{0i}) \frac{\pi_{0i}}{1 - \pi_{0i}}}. \qquad (15.27)$$

Infine, per quanto riguarda le probabilità di inclusione del secondo ordine, usando la (15.27) è immediato verificare, con ovvia simbologia, che

$$\begin{aligned} \pi_{ij} &= p(\{i, j\}) \\ &= \frac{\pi_{0i}\, \pi_{0j}}{\sum_{i=1}^{N} (2 - \pi_{0i}) \frac{\pi_{0i}}{1 - \pi_{0i}}} \left\{ \frac{1}{1 - \pi_{0i}} + \frac{1}{1 - \pi_{0j}} \right\}. \end{aligned}$$

È facile verificare (Esercizio 15.9) che vale la relazione $\pi_{ij} \leqslant \pi_i \, \pi_j$ per tutte le coppie di unità distinte della popolazione. □

Giova sottolineare, in chiusura, che il calcolo delle probabilità di inclusione del secondo ordine è molto semplificato se si usa l'espressione approssimata π^a_{ij} sviluppata nella Sezione 12.7, la quale fornisce risultati molto buoni nel caso dello schema di Sampford.

Esempio 15.2. Si considerino una popolazione di $N = 10$ unità, e una numerosità campionaria $n = 5$. Le probabilità di inclusione del primo ordine desiderate, esattamente come in Sampford (1967) e Hájek (1981), p. 90, sono: $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$. Chiaramente, si ha:

$$d = \sum_{i=1}^{10} \pi_{0i}(1 - \pi_{0i}) = 2.09.$$

Per approssimare le probabilità di inclusione del secondo ordine è necessario calcolare le quantità c_i, le quali risultano pari a:

$$c_1 = 0.0613,\ c_2 = 0.1525,\ c_3 = 0.1670,\ c_4 = 0.1840,\ c_5 = 0.1861,$$
$$c_6 = 0.1861,\ c_7 = 0.1776,\ c_8 = 0.1670,\ c_9 = 0.1344,\ c_{10} = 0.1130.$$

Tabella 15.1 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – disegno di Sampford

0.9000	0.6212	0.5752	0.4840	0.4387	0.4387	0.3489	0.3044	0.2163	0.1726
0.6212	0.7000	0.4303	0.3573	0.3217	0.3217	0.2526	0.2191	0.1539	0.1222
0.5752	0.4303	0.6500	0.3271	0.2940	0.2940	0.2301	0.1992	0.1396	0.1106
0.4840	0.3573	0.3271	0.5500	0.2407	0.2407	0.1871	0.1616	0.1126	0.0890
0.4387	0.3217	0.2940	0.2407	0.5000	0.2152	0.1668	0.1438	0.1000	0.0790
0.4387	0.3217	0.2940	0.2407	0.2152	0.5000	0.1668	0.1438	0.1000	0.0790
0.3489	0.2526	0.2301	0.1871	0.1668	0.1668	0.4000	0.1106	0.0766	0.0604
0.3044	0.2191	0.1992	0.1616	0.1438	0.1438	0.1106	0.3500	0.0658	0.0517
0.2163	0.1539	0.1396	0.1126	0.1000	0.1000	0.0766	0.0658	0.2500	0.0355
0.1726	0.1222	0.1106	0.0890	0.0790	0.0790	0.0604	0.0517	0.0355	0.2000

Tabella 15.2 Valori approssimati π^a_{ij} delle probabilità di inclusione del secondo ordine

0.9000	0.6206	0.5748	0.4837	0.4386	0.4386	0.3491	0.3048	0.2168	0.1730
0.6206	0.7000	0.4295	0.3569	0.3216	0.3216	0.2529	0.2195	0.1545	0.1228
0.5748	0.4295	0.6500	0.3268	0.2939	0.2939	0.2303	0.1996	0.1400	0.1111
0.4837	0.3569	0.3268	0.5500	0.2408	0.2408	0.1873	0.1618	0.1128	0.0892
0.4386	0.3216	0.2939	0.2408	0.5000	0.2153	0.1669	0.1439	0.1000	0.0790
0.4386	0.3216	0.2939	0.2408	0.2153	0.5000	0.1669	0.1439	0.1000	0.0790
0.3491	0.2529	0.2303	0.1873	0.1669	0.1669	0.4000	0.1103	0.0761	0.0599
0.3048	0.2195	0.1996	0.1618	0.1439	0.1439	0.1103	0.3500	0.0650	0.0511
0.2168	0.1545	0.1400	0.1128	0.1000	0.1000	0.0761	0.0650	0.2500	0.0348
0.1730	0.1228	0.1111	0.0892	0.0790	0.0790	0.0599	0.0511	0.0348	0.2000

Nella Tabella 15.1 sono riportate le probabilità di inclusione del secondo ordine esatte, mentre nella Tabella 15.2 sono riportate le probabilità di inclusione del secondo ordine approssimate. Benché la numerosità della popolazione (e di conseguenza il valore di d) non siano elevati, i valori approssimati π_{ij}^a sono assai vicini a quelli effettivi π_{ij}. □

15.3.2 Implementazione del disegno di Sampford

Esistono diversi algoritmi (schemi) per l'implementazione del disegno di Sampford. Il più semplice è sicuramente lo *schema multinomiale* (uno schema di accettazione condizionata, in sostanza) di seguito esposto.

- *Passo* 0. Inizializzazione. Porre $m = 0$, $B_1 = 0$, ..., $B_N = 0$.
- *Passo* 1. Scegliere un'unità della popolazione, in maniera tale che l'unità i abbia probabilità p_i/n di essere selezionata. Se si sceglie l'unità i, porre $B_i = 1$. Incrementare m di 1, e andare al Passo 2.
- *Passo* 2. Se $m = n$ andare al Passo 4. Se invece $m < n$, scegliere un'unità dalla popolazione in modo tale che l'unità i abbia probabilità

$$\mathrm{P}_i = \frac{\frac{p_i}{1-p_i}}{\sum_{j=1}^{N} \frac{p_j}{1-p_j}}; \quad i = 1, \ldots, N$$

  di essere selezionata. Andare al Passo 3.
- *Passo* 3. Se si è scelta l'unità i ed è $B_i = 1$, andare al Passo 0.
  Altrimenti, porre $B_i = 1$, incrementare m di 1, e andare al Passo 2.
- *Passo* 4. Arresto. Il campione s è formato dalle n unità i tali che $B_i = 1$.

Uno schema alternativo di implementazione del disegno di Sampford, basato su un algoritmo di rigetto di campioni di Pareto, è sviluppato nell'Esercizio 15.13.

15.4 Il disegno campionario di tipo Pareto

Il disegno campionario di tipo Pareto, introdotto in Rosén (1997a), Rosén (1997b), pur se possiede solo in via approssimata i "requisiti desiderabili" elencati nella Sezione 15.1, ha un'importante proprietà: è estremamente semplice da implementare. Per questa ragione, e per il fatto che è usato come "schema accessorio" per implementare anche altri disegni campionari, esso viene esposto nella presente sezione.

15.4.1 Aspetti essenziali di base

Si considerino N numeri positivi $\lambda_1, \ldots, \lambda_N$, tali che $0 < \lambda_i < 1$ e

$$\sum_{i=1}^{N} \lambda_i = n.$$

La selezione di un campione di numerosità n mediante *disegno di Pareto* consiste dei seguenti passi.

- *Passo* 1. Si generano N variabili aletaorie $U_1, \ldots, U_N$ indipendenti, ciascuna con distribuzione uniforme in $[0, 1]$.
- *Passo* 2. Si calcolano le quantità

$$Q_i = \frac{\frac{U_i}{1-U_i}}{\frac{\lambda_i}{1-\lambda_i}}, \quad i = 1, \ldots, N.$$

- *Passo* 3. Si ordinano le quantità $Q_1, \ldots, Q_N$ dalla più piccola alla più grande. Per convenzione, indicheremo $Q_{1:N}$ la più piccola tra le Q_i, con $Q_{2:N}$ la seconda più piccola tra le Q_i, e così via, fino a $Q_{N:N}$ che è la più grande tra le Q_i. Si ha quindi

$$Q_{1:N} < Q_{2:N} < \cdots < Q_{N:N}.$$

- *Passo* 4. Il campione s è formato dalle n unità con i valori Q_i più piccoli. Formalmente, l'unità i entra a far parte del campione s se il corrispondente Q_i è uno tra $Q_{1:N}$, $Q_{2:N}$, ..., $Q_{n:N}$.

I passi 1-4 chiariscono bene un punto importante: selezionare un campione s con lo schema di tipo Pareto è estremamente semplice. Due sono, però, i problemi da risolvere.

Pa 1. Dati gli n numeri $\lambda_1, \ldots, \lambda_N$, quali valori assumono le probabilità di inclusione del primo ordine $\pi_1, \ldots, \pi_N$? Lo stesso quesito vale per le probabilità di inclusione del secondo ordine.

Pa 2. Quali valori devono assumere $\lambda_1, \ldots, \lambda_N$ in modo che le probabilità di inclusione del primo ordine assumano i valori desiderati $\pi_{01}, \ldots, \pi_{0N}$?

I due quesiti sono ovviamente legati, nel senso che rispondere all'uno porta implicitamente a rispondere all'altro. Qui ci accontenteremo di dare soltanto una risposta approssimata. Qualche approfondimento sul calcolo esatto delle probabilità di inclusione sarà fornito più avanti.

Incominciamo a rispondere al quesito *Pa 2*. Date le probabilità di inclusione desiderate, $\pi_{01}, \ldots, \pi_{0N}$, come mostrato in Bondesson *e altri* (2006), una buona approssimazione per i valori λ_i consiste nel porre:

$$\frac{\lambda_i}{1-\lambda_i} = c\,\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{-\frac{1}{d^2}\,\pi_{0i}\,(1-\pi_{0i})\,\left(\pi_{0i}-\frac{1}{2}\right)\right\}, \quad i = 1, \ldots, N \tag{15.28}$$

essendo c una costante di proporzionalità, tale che $\lambda_1 + \cdots + \lambda_N = n$, e

$$d = \sum_{i=1}^{N} \pi_{0i}(1 - \pi_{0i}).$$

Il valore che deve assumere la costante c nella (15.28) non è in generale determinabile esplicitamente, ma va calcolato per via numerica. Tuttavia, l'effettiva determinazione di c è sostanzialmente inutile. Infatti, essendo c una costante positiva, dividere le variabili $U_i/(1 - U_i)$ per i termini

$$c \frac{\pi_i}{1 - \pi_i} \exp\left\{-\frac{1}{d^2}\,\pi_{0i}\,(1 - \pi_{0i})\,\left(\pi_{0i} - \frac{1}{2}\right)\right\}$$

o dividerle per

$$\frac{\pi_{0i}}{1 - \pi_{0i}} \exp\left\{-\frac{1}{d^2}\,\pi_{0i}\,(1 - \pi_{0i})\,\left(\pi_{0i} - \frac{1}{2}\right)\right\}$$

non cambia il loro ordine. In altre parole, nel Passo 2 si può porre

$$Q_i = \frac{\frac{U_i}{1-U_i}}{\frac{\pi_{0i}}{1-\pi_{0i}} \exp\left\{-\frac{1}{d^2}\,\pi_{0i}\,(1 - \pi_{0i})\,\left(\pi_{0i} - \frac{1}{2}\right)\right\}}, \quad i = 1, \ldots, N$$

e procedere come indicato nei passi successivi.

Una semplificazione si ottiene per valori "grandi" di d. In tal caso si ha infatti $1/d^2 \approx 0$, da cui

$$\exp\left\{-\frac{1}{d^2}\,\pi_{0i}\,(1 - \pi_{0i})\,\left(\pi_{0i} - \frac{1}{2}\right)\right\} \approx 1$$

e quindi si può scrivere $\lambda_i \approx \pi_{0i}$.

Per quanto concerne, infine, le probabilità di inclusione del secondo ordine, risultati abbastanza soddisfacenti si ottengono tramite l'approssimazione sviluppata nella Sezione 12.7.

Un caso molto speciale è quello in cui $\lambda_1 = \lambda_2 = \cdots = \lambda_N$. È infatti facile verificare (Esercizio 15.10) che in questo caso il disegno di tipo Pareto si riduce a quello semplice senza ripetizione.

Esempio 15.3. Si consideri la popolazione di $N = 10$ unità dell'Esempio 15.2, in cui la numerosità campionaria è $n = 5$ e le probabilità di inclusione del primo ordine desiderate sono $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$. Come già visto, si ha $d = \sum_i \pi_{0i}(1 - \pi_{0i}) = 2.09$.

Le quantità al membro di destra della (15.28), a meno della costante c, risultano nel caso in esame pari ai valori riportati in Tabella 15.3.

Le corrispondenti probabilità di inclusione del primo ordine *effettive*, calcolate numericamente, sono riportate in Tabella 15.4. Come si vede, si tratta di valori vicinissimi a quelli delle probabilità di inclusione desiderate.

Nella Tabella 15.5 sono riportate le probabilità di inclusione del secondo ordine *effettive*, sempre calcolate per via numerica.

I valori approssimati π_{ij}^a delle probabilità di inclusione del secondo ordine, ottenuti usando l'approccio delineato nella Sezione 12.7, sono riportati in Tabella 15.2. Come si può osservare, l'approssimazione $\pi_{ij} \approx \pi_{ij}^a$ è abbastanza soddisfacente. □

Tabella 15.3 Valori di $\lambda_i/(1 - \lambda_i)$ calcolati con la (15.28)

i	1	2	3	4	5	6	7	8	9	10
$\frac{\lambda_i}{1-\lambda_i}$	8.926	2.311	1.843	1.219	1.00 0	1.000	0.670	0.543	0.337	0.253

Tabella 15.4 Probabilità di inclusione effettive del primo ordine

i	1	2	3	4	5	6	7	8	9	10
π_i	0.9001	0.7002	0.6502	0.5502	0.5003	0.5503	0.3995	0.3498	0.2497	0.1997

Tabella 15.5 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – disegno di Pareto

0.9001	0.6215	0.5756	0.4843	0.4390	0.4390	0.3487	0.3043	0.2160	0.1724
0.6215	0.7002	0.4307	0.3575	0.3220	0.3218	0.2523	0.2191	0.1537	0.1220
0.5756	0.4307	0.6502	0.3273	0.2943	0.2942	0.2298	0.1991	0.1395	0.1104
0.4843	0.3575	0.3273	0.5502	0.2409	0.2410	0.1870	0.1616	0.1125	0.0888
0.4390	0.3220	0.2943	0.2409	0.5003	0.2153	0.1667	0.1438	0.1000	0.0790
0.4390	0.3218	0.2942	0.2410	0.2153	0.5003	0.1665	0.1438	0.0997	0.0789
0.3487	0.2523	0.2298	0.1870	0.1667	0.1665	0.3995	0.1105	0.0764	0.0603
0.3043	0.2191	0.1991	0.1616	0.1438	0.1438	0.1105	0.3498	0.0656	0.0515
0.2160	0.1537	0.1395	0.1125	0.1000	0.0997	0.0764	0.0656	0.2497	0.0354
0.1724	0.1220	0.1104	0.0888	0.0790	0.0789	0.0603	0.0515	0.0354	0.1997

*15.4.2 Approfondimenti: probabilità dei campioni nel disegno di Pareto**

Lo spazio dei campioni (di unità) nel disegno di Pareto è ovviamente l'insieme $\mathcal{C}_{N,n}$ di tutte le combinazioni senza ripetizione di classe n delle unità della popolazione. Per quanto riguarda la determinazione delle probabilità dei campioni, iniziamo con l'osservare che se $\mathbf{s}$ è il campione di cui si vuole determinare la probabilità, allora una delle unità i di $\mathbf{s}$ deve corrispondere all'n-mo valore Q ordinato, $Q_{n:N}$, mentre le altre $n-1$ unità possiedono valori Q più piccoli. In altre parole, vale la seguente equivalenza logica:

Il campione selezionato è $\mathbf{s}$ se e solo se:

$$\begin{aligned} &Q_i = Q_{n:N} \textit{ per un'unità } i \in \mathbf{s}, \text{ e} \\ &Q_j < Q_i \textit{ per le altre unità } \in \mathbf{s},\ j \neq i, \text{ e} \\ &Q_k > Q_i \textit{ per tutte le unità } k \notin \mathbf{s}. \end{aligned}$$

Si può quindi scrivere

$$p(\mathbf{s}) = \sum_{i \in \mathbf{s}} Pr(Q_i = Q_{n:N};\ Q_j < Q_i \textit{ per } j \in \mathbf{s},\ j \neq i;\ Q_k > Q_i \textit{ per } k \notin \mathbf{s}). \quad (15.29)$$

La determinazione effettiva della (15.29) richiede qualche complicazione aggiuntiva. In primo luogo, detta f_{Q_i} la funzione di densità della variabile aleatoria Q_i, dalla (15.29) si evince subito che

$$\begin{aligned} p(\mathbf{s}) &= \sum_{i \in \mathbf{s}} \int_{-\infty}^{+\infty} Pr(Q_j < y \textit{ per } j \in \mathbf{s},\ j \neq i;\ Q_k > y \textit{ per } k \notin \mathbf{s})\, f_{Q_i}(y)\, dy \\ &= \sum_{i \in \mathbf{s}} \int_{-\infty}^{+\infty} Pr(Q_j < y \textit{ per } j \in \mathbf{s},\ j \neq i)\, Pr(Q_k > y \textit{ per } k \notin \mathbf{s})\, f_{Q_i}(y)\, dy \\ &= \sum_{i \in \mathbf{s}} \int_{-\infty}^{+\infty} \left\{ \prod_{\substack{j \in \mathbf{s} \\ j \neq i}} Pr(Q_j < y) \right\} \left\{ \prod_{k \notin \mathbf{s}} Pr(Q_k > y) \right\} f_{Q_i}(y)\, dy. \end{aligned} \quad (15.30)$$

Posto

$$\vartheta_i = \frac{\lambda_i}{1 - \lambda_i}, \quad \textit{per ciascuna unità } i = 1, \ldots, N \quad (15.31)$$

è facile verificare (Esercizio 15.11) che

$$Pr(Q_i \leqslant y) = \frac{\vartheta_i\, y}{1 + \vartheta_i\, y}, \quad y \geqslant 0 \quad (15.32)$$

da cui discende che $f_{Q_i}(y) = \vartheta_i/(1+\vartheta_i\, y)^2$ (per $y \geqslant 0$), e quindi, usando la (15.30)

$$
\begin{aligned}
p(\mathbf{s}) &= \sum_{i\in\mathbf{s}} \int_0^\infty \left\{ \prod_{\substack{j\in\mathbf{s}\\ j\neq i}} \frac{\vartheta_j\, y}{1+\vartheta_j\, y} \right\} \left\{ \prod_{k\notin\mathbf{s}} \frac{1}{1+\vartheta_k\, y} \right\} \frac{\vartheta_i}{(1+\vartheta_i\, y)^2}\, dy \\
&= \left\{ \prod_{j\in\mathbf{s}} \vartheta_j \right\} \sum_{i\in\mathbf{s}} \int_0^\infty \left\{ \prod_{k=1}^N \frac{1}{1+\vartheta_k\, y} \right\} \frac{y^{n-1}}{1+\vartheta_i\, y}\, dy \\
&= \left\{ \prod_{j=1}^N \vartheta_j^{\delta(j;\mathbf{s})} \right\} \left\{ \sum_{i=1}^N \delta(i;\, \mathbf{s}) \int_0^\infty \left(\prod_{k=1}^N \frac{1}{1+\vartheta_k\, y} \right) \frac{y^{n-1}}{1+\vartheta_i\, y}\, dy \right\} \\
&= C_{pa} \left\{ \prod_{j=1}^N \vartheta_j^{\delta(j;\mathbf{s})} \right\} \left\{ \sum_{i=1}^N \delta(i;\, \mathbf{s}) \int_0^\infty \left(\prod_{k=1}^N \frac{1+\vartheta_k}{1+\vartheta_k\, y} \right) \frac{y^{n-1}}{1+\vartheta_i\, y}\, dy \right\} \\
&= \left(\prod_{j=1}^N \lambda_j^{\delta(j;\mathbf{s})} (1-\lambda_j)^{1-\delta(j;\mathbf{s})} \right) \left\{ \sum_{i=1}^N g_i\, \delta(i;\, \mathbf{s}) \right\} \qquad (15.33)
\end{aligned}
$$

con

$$
g_i = \int_0^\infty \left(\prod_{k=1}^N \frac{1+\vartheta_k}{1+\vartheta_k\, y} \right) \frac{y^{n-1}}{1+\vartheta_i\, y}\, dy; \quad i = 1, \ldots, N. \qquad (15.34)
$$

La (15.33) si può anche scrivere come

$$
p(\mathbf{s}) = C_{pa} \left\{ \prod_{j=1}^N \vartheta_j^{\delta(j;\mathbf{s})} \right\} \left\{ \sum_{i=1}^N g_i\, \delta(i;\, \mathbf{s}) \right\} \qquad (15.35)
$$

avendo posto

$$
C_{pa} = \prod_{k=1}^N \frac{1}{1+\vartheta_k}.
$$

Le costanti $g_1, \ldots, g_N$ si possono ricavare esplicitamente per via analitica, benché la loro espressione sia complicata e non molto adatta al calcolo numerico. Piú semplice, e tutto sommato più conveniente, è il loro calcolo per via numerica. In alternativa, un'espressione approssimata dovuta a Bondesson

e altri (2006), che in genere produce buoni risultati, è la seguente:

$$g_i \approx (1 - \lambda_i)\,\sqrt{2\,\pi}\,\gamma_i\,\exp\left\{\frac{1}{2}\,\gamma_i^2\,\lambda_i^2\right\} \tag{15.36}$$

dove si è posto

$$\gamma_i^2 = \frac{1}{L + \lambda_i(1 - \lambda_i)}, \quad i = 1, \ldots, N$$
$$L = \sum_{i=1}^{N} \lambda_i\,(1 - \lambda_i).$$

Si osservi che per L "grande" si ha $\gamma_i^2 \approx 0$, e quindi $g_i \approx 1 - \lambda_i$.

Dalla (15.35) sarebbe in linea di principio possibile, benché computazionalmente difficile, calcolare le probabilità di inclusione esatte del primo e del secondo ordine. Qualche cenno in proposito è nell'Esercizio 15.13. Per approfondimenti si rinvia al lavoro di Bondesson (2010).

15.5 Il disegno campionario di Poisson condizionato

15.5.1 Aspetti introduttivi e di base

Il disegno di Poisson condizionato nasce essenzialmente come "correzione" del maggior difetto del disegno di Poisson: la numerosità campionaria variabile. In questa sezione ci limiteremo ad un'esposizione molto succinta del disegno di Poisson condizionato. Per una trattazione approfondita si rinvia al Capitolo 7 del volume di Hájek (1981).

Si considerino N numeri $\tau_1, \ldots, \tau_N$, tali che

$$0 < \tau_i < 1 \;\; \textit{per ciascuna unità } i = 1, \ldots, N.$$

Il disegno di Poisson condizionato è definito dalle seguenti due specificazioni.

- lo spazio dei campioni è l'insieme $\mathcal{C}_{N,n}$ di tutte le combinazioni senza ripetizioni di classe n;
- ciascun campione $\mathbf{s} \in \mathcal{C}_{N,n}$ ha probabilità

$$p(\mathbf{s}) = A_{pc} \prod_{i=1}^{N} \tau_i^{\delta(i;\,\mathbf{s})} (1 - \tau_i)^{1 - \delta(i;\,\mathbf{s})} \tag{15.37}$$

dove A_{ps} è una costante opportuna, tale che la somma delle (15.37) sia uguale a 1.

Se, in modo simile a quanto fatto in precedenza, si definiscono i numeri

$$\eta_i = \frac{\tau_i}{1 - \tau_i} \;\; \textit{per ciascuna unità } i = 1, \ldots, N \tag{15.38}$$

la (15.37) si riscrive in forma equivalente come

$$p(\mathbf{s}) = C_{pc} \prod_{i=1}^{N} \eta_i^{\delta(i;\mathbf{s})} \tag{15.39}$$

essendo la costante C_{pc} tale da rendere pari a 1 la somma delle (15.39). A meno di casi molto speciali, essa non è calcolabile esplicitamente.

Un disegno di Poisson condizionato è *normalizzato* se:

$$\tau_1 + \tau_2 + \cdots + \tau_N = n.$$

Il termine "disegno di Poisson condizionato" deriva dal fatto che se $\mathbf{s}$ è un campione di n unità selezionato mediante il disegno di Poisson introdotto in precedenza, si ha

$$\begin{aligned} p(\mathbf{s}\,|n(\mathbf{s}) = n) &= \frac{p(\mathbf{s})}{Pr(n(\mathbf{s}) = n)} \\ &= \frac{C_{po} \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{\delta(i;\mathbf{s})}}{C_{po} \sum_{\mathbf{s}\in\mathcal{C}_{N,n}} \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{\delta(i;\mathbf{s})}} = \frac{\prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{\delta(i;\mathbf{s})}}{\sum_{\mathbf{s}\in\mathcal{C}_{N,n}} \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{\delta(i;\mathbf{s})}} \\ &= \frac{\prod_{i=1}^{N} \omega_i^{\delta(i;\mathbf{s})}}{\sum_{\mathbf{s}\in\mathcal{C}_{N,n}} \prod_{i=1}^{N} \omega_i^{\delta(i;\mathbf{s})}} \quad \textit{per ciascun campione } \mathbf{s} \in \mathcal{C}_{N,n} \end{aligned}$$

ossia proprio la (15.39), con p_i al posto di τ_i e, di conseguenza, ω_i in luogo di η_i.

Una proprietà rimarchevole del disegno di Poisson condizionato riguarda la sua entropia. Come conseguenza del suo legame con il disegno di Poisson "non condizionato", non è difficile verificare che tra tutti i disegni campionari non ordinati, senza ripetizioni, ad ampiezza effettiva costante n e con prefissate probabilità di inclusione del primo ordine, il disegno di Poisson condizionato è quello di entropia massima.

Benché concettualmente semplice, il disegno di Poisson condizionato presenta diversi aspetti che vanno chiariti, per renderne possibile l'applicazione. Essendo in generale $\pi_i \neq \tau_i$ (al contrario di quel che accade per i disegni di Sampford e di Poisson), il primo punto critico riguarda il calcolo delle probabilità di inclusione. Il secondo punto, poi, riguarda l'implementazione del disegno di Poisson condizionato, ovvero la costruzione di uno schema numericamente efficiente per la selezione di un campione.

Per quanto riguarda le probabilità di inclusione, due sono, al solito, i problemi da risolvere.

Pc 1. Dati gli n numeri $\tau_1, \ldots, \tau_N$ quali valori assumono le probabilità di inclusione del primo ordine $\pi_1, \ldots, \pi_N$? Lo stesso discorso vale, ovviamente, per le probabilità di inclusione del secondo ordine.

Pc 2. Quali valori devono assumere $\tau_1, \ldots, \tau_N$ in modo che le probabilità di inclusione del primo ordine assumano i valori desiderati $\pi_{01}, \ldots, \pi_{0N}$?

I due quesiti sono ovviamente legati, nel senso che rispondere all'uno porta implicitamente a rispondere all'altro. Qui ci accontenteremo di dare soltanto una risposta approssimata.

Probabilità di inclusione: calcolo approssimato

In questa parte daremo una semplice approssimazione delle probabilità di inclusione del primo ordine del disegno di Poisson, che ci consentirà di fornire una soluzione semplice, nell'ordine, ai problemi *Pc* 2 e *Pc* 1. Dette, come sempre, $\pi_{01}, \ldots, \pi_{0N}$ le probabilità di inclusione del primo ordine desiderate, poniamo

$$d = \sum_{i=1}^{N} \pi_{0i}(1 - \pi_{0i}). \tag{15.40}$$

L'idea di base è di prendere i numeri τ_i in modo che sia soddisfatta la relazione

$$\frac{\tau_i}{1-\tau_i} = c\,\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\} \quad \textit{per ciascuna unità } i = 1 \ldots, N \tag{15.41}$$

essendo c una costante tale che sia $0 < \tau_i \leqslant 1$ e $\tau_1 + \cdots + \tau_N = n$. Con questa scelta dei numeri τ_i, e tenendo presente la (15.38), la (15.39) diventa

$$\begin{aligned} p(\mathbf{s}) &= C_{pc}\left\{\prod_{i=1}^{N}\left(\frac{\pi_{0i}}{1-\pi_{0i}}\right)^{\delta(i;\,\mathbf{s})}\right\}\exp\left\{\frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,\delta(i;\,\mathbf{s})\right\} \\ &= \widetilde{C}_{pc}\left\{\prod_{i=1}^{N}\left(\frac{\pi_{0i}}{1-\pi_{0i}}\right)^{\delta(i;\,\mathbf{s})}\right\}\exp\left\{\frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,(\delta(i;\,\mathbf{s})-\pi_{0i})\right\} \end{aligned}$$

con $\widetilde{C}_{pc}$ costante di proporzionalità opportuna. D'altra parte, con un semplice sviluppo di Taylor si ha

$$\begin{aligned} \exp\left\{\frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,(\delta(i;\,\mathbf{s})-\pi_{0i})\right\} &= 1 + \frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,(\delta(i;\,\mathbf{s})-\pi_{0i}) + \cdots \\ &\approx 1 + \frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,(\delta(i;\,\mathbf{s})-\pi_{0i}) \\ &= \frac{1}{d}\sum_{i=1}^{N}(1-\pi_{0i})\,\delta(i;\,\mathbf{s}). \end{aligned}$$

Pertanto, la scelta (15.41) produce un disegno di Poisson condizionato in cui le probabilità dei campioni sono *in via approssimata* proporzionali a

$$\left\{\prod_{i=1}^{N}\left(\frac{\pi_{0i}}{1-\pi_{0i}}\right)^{\delta(i;\,\mathbf{s})}\right\}\sum_{i=1}^{N}(1-\pi_{0i})\,\delta(i;\,\mathbf{s})\,.$$

D'altra parte, la (15.18) mostra che le probabilità dei campioni del disegno di Sampford, con $p_i = \pi_{0i}$ sono *esattamente* proporzionali agli stessi fattori. Pertanto, con la scelta (15.41) il disegno di Poisson condizionato può essere approssimato con un disegno di Sampford in cui $p_i = \pi_{0i}$. Poiché quest'ultimo ha probabilità di inclusione esattamente uguali alle π_{0i}, si conclude che lo scegliere valori τ_i che soddisfano la (15.41) produce un disegno di Poisson in cui le probabilità di inclusione del primo ordine sono *approssimativamente* uguali ai valori desiderati $\pi_{01}, \ldots, \pi_{0N}$.

Per determinare un po' più esplicitamente i numeri τ_i sulla base della (15.41), iniziamo con l'osservare che tale relazione equivale a

$$\tau_i = c\,(1-\tau_i)\,\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\}$$

da cui si ottiene

$$\tau_i = \frac{\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\}}{\frac{1}{c}+\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\}} \quad \textit{per ciascuna unità } i = 1, \ldots, N. \tag{15.42}$$

La costante c si ricava in modo che sia soddisfatta la relazione $\tau_1 + \cdots + \tau_N = n$, che nel nostro caso si riscrive come

$$\sum_{i=1}^{N} \frac{\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\}}{\frac{1}{c}+\frac{\pi_{0i}}{1-\pi_{0i}}\,\exp\left\{\frac{1-\pi_{0i}}{d}\right\}} = n. \tag{15.43}$$

Purtroppo quest'equazione non è risolvibile per via analitica. Occorre utilizzare un qualche metodo numerico, come ad esempio il *metodo delle bisezioni*, che si applica in maniera simile a quanto illustrato nella Sezione 12.7.

Quanto sopra esposto fornisce non solo una soluzione approssimata al problema di determinare i numeri τ_i in modo che il disegno di Poisson condizionato abbia le desiderate probabilità di inclusione π_{0i}, ma anche, sia pur sempre in via approssimata, al problema $Pc\,2$. Dati $\tau_1, \ldots, \tau_N$, invertendo le relazioni (15.41) si ottiene

$$\frac{\pi_i}{1-\pi_i} = c'\,\frac{\tau_i}{1-\tau_i}\,\exp\left\{-\frac{1-\tau_i}{T}\right\} \quad \textit{per ciascuna unità } i = 1, \ldots, N$$

essendo c' un'opportuna costante tale che $\pi_1 + \cdots + \pi_N = n$, essendo $T = \sum_i \tau_i(1-\tau_i)$. La determinazione effettiva delle probabilità di inclusione π_i, che richiede anche il calcolo esplicito della costante c', si può effettuare seguendo le stesse linee indicate per il calcolo dei termini τ_i dati i valori $\pi_{01}, \ldots, \pi_{0N}$.

Se il termine $d = \sum_i \pi_{0i}(1-\pi_{0i})$ è "grande", si ha $(1-\pi_{0i})/d \approx 0$, e quindi

$$\exp\left\{\frac{1-\pi_{0i}}{d}\right\} \approx 1.$$

In questo caso la (15.41) fornisce la semplicissima relazione $\tau_i = \pi_{0i}$.

Le idee di base per l'approssimazione (15.42) sono in Bondesson *e altri* (2006). Per un approccio differente, che conduce ad una diversa approssimazione (che comunque, per d grande, si riduce a $\tau_i = \pi_{0i}$), si rinvia al volume di Hájek (1981), p. 72.

Per quanto riguarda le probabilità di inclusione del secondo ordine, un'approssimazione che nel caso del disegno di Poisson condizionato fornisce buoni risultati è quella sviluppata nella Sezione 12.7.

Esempio 15.4. Si consideri la popolazione di $N = 10$ unità dell'Esempio 15.2, in cui la numerosità campionaria è $n = 5$ e le probabilità di inclusione del primo ordine desiderate sono $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$. Si ha $d = \sum_i \pi_{0i}(1 - \pi_{0i}) = 2.09$.

I valori dei coefficienti τ_i, calcolati in base alla (15.42), sono riportati in Tabella 15.6.

Le probabilità di inclusione del primo ordine *effettive* sono riportate in Tabella 15.7. I loro valori sono pressoché identici a quelli desiderati.

Nella Tabella 15.8 sono riportate le probabilità di inclusione del secondo ordine *effettive*, calcolate per via numerica.

Tabella 15.6 Valori di τ_i calcolati con la (15.42)

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
τ_i	0.880	0.679	0.633	0.544	0.500	0.500	0.411	0.366	0.273	0.224

Tabella 15.7 Probabilità di inclusione effettive del primo ordine

i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
π_i	0.8989	0.6998	0.6499	0.5501	0.5000	0.5500	0.4000	0.3500	0.2506	0.2007

Tabella 15.8 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – disegno di Poisson condizionato

0.8989	0.6205	0.5747	0.4836	0.4380	0.4380	0.3483	0.3038	0.2160	0.1725
0.6205	0.6998	0.4306	0.3576	0.3218	0.3218	0.2523	0.2187	0.1538	0.1220
0.5747	0.4306	0.6499	0.3273	0.2940	0.2940	0.2300	0.1990	0.1394	0.1107
0.4836	0.3576	0.3273	0.5501	0.2404	0.2404	0.1873	0.1614	0.1128	0.0894
0.4379	0.3218	0.2941	0.2407	0.5000	0.2150	0.1669	0.1438	0.1002	0.0793
0.4381	0.3218	0.2941	0.2407	0.2150	0.5000	0.1669	0.1438	0.1002	0.0793
0.3483	0.2523	0.2230	0.1873	0.1669	0.1669	0.4000	0.1109	0.0771	0.0609
0.3038	0.2187	0.1990	0.1614	0.1438	0.1438	0.1109	0.3500	0.0664	0.0523
0.2160	0.1538	0.1394	0.1128	0.1002	0.1002	0.0771	0.0664	0.2506	0.0363
0.1725	0.1220	0.1107	0.0894	0.0794	0.0794	0.0609	0.0523	0.0363	0.2007

I valori approssimati π_{ij}^a delle probabilità di inclusione del secondo ordine, ottenuti usando l'approccio delineato nella Sezione 12.7, sono quelli della Tabella 15.2. L'approssimazione $\pi_{ij} \approx \pi_{ij}^a$ è soddisfacente. □

Probabilità di inclusione: calcolo esatto*

Può essere di interesse anche notevole, soprattutto per valori moderati o piccoli di $d = \sum_i \pi_{0i}(1 - \pi_{0i})$, cercare di calcolare i valori esatti delle probabilità di inclusione del primo e del secondo ordine. In questa sezione svilupperemo alcune semplici relazioni ricorsive, che si prestano ad un'efficiente calcolo numerico delle probabilità di inclusione. La tecnica e il simbolismo sono abbastanza simili a quelli usati per studiare le proprietà del disegno di Sampford. Per varianti e approfondimenti, si rinvia al volume di Tillé (2006), pp. 79–88, e all'articolo Bondesson (2010).

Dati gli N numeri $\tau_1, \ldots, \tau_N$, tutti compresi tra 0 e 1, consideriamo il disegno di Poisson condizionato di numerosità n da essi individuato. Rispetto al simbolismo usato nelle precedenti sezioni è necessario introdurre una (lieve) complicazione nella notazione, dovuta al fatto che è necessario inserire esplicitamente nei simboli la numerosità campionaria. Detto $\mathbf{s}_n$ un generico campione (combinazione senza ripetizione di n unità della popolazione), la sua probabilità di selezione, come risulta dalla (15.39) è pari a

$$p^{(n)}(\mathbf{s}_n) = C_{pc}^{(n)} \prod_{k=1}^{N} \eta_k^{\delta(k;\,\mathbf{s}_n)}.$$

La probabilità di inclusione del primo ordine dell'unità i, indicata con $\pi_i^{(n)}$, è eguale a

$$\pi_i^{(n)} = \sum_{\mathbf{s}_n \in \mathcal{C}_{N,n}} \delta(i;\,\mathbf{s}_n)\, p^{(n)}(\mathbf{s}_n). \tag{15.44}$$

Parallelamente, consideriamo il disegno di Poisson condizionato, sempre definito da $\tau_1, \ldots, \tau_N$, ma di numerosità campionaria $n - 1$. Detto $\mathbf{s}_{n-1}$ un generico campione (combinazione senza ripetizione di $n - 1$ unità della popolazione), la sua probabilità di selezione è

$$p^{(n-1)}(\mathbf{s}_{n-1}) = C_{pc}^{(n-1)} \prod_{k=1}^{N} \eta_k^{\delta(k;\,\mathbf{s}_{n-1})}.$$

La probabilità di inclusione del primo ordine dell'unità i, indicata con $\pi_i^{(n-1)}$, è invece

$$\pi_i^{(n-1)} = \sum_{\mathbf{s}_{n-1} \in \mathcal{C}_{N,n-1}} \delta(i;\,\mathbf{s}_{n-1})\, p^{(n-1)}(\mathbf{s}_{n-1}). \tag{15.45}$$

Non è difficile verificare (Esercizio 15.12) che vale la relazione:

$$\pi_i^{(n)} = \frac{C_{pc}^{(n)}}{C_{pc}^{(n-1)}}\,\eta_i\,\left(1-\pi_i^{(n-1)}\right) \quad \textit{per ciascuna unità } i=1,\,\ldots\,N. \tag{15.46}$$

Per sfruttare appieno la (15.46) occorre determinare il rapporto $C_{pc}^{(n)}/C_{pc}^{(n-1)}$. Dalla (15.46) discende che

$$n=\sum_{i=1}^{N}\pi_i^{(n)}=\frac{C_{pc}^{(n)}}{C_{pc}^{(n-1)}}\sum_{j=1}^{N}\eta_j\,(1-\pi_j^{(n-1)})$$

da cui si ottiene

$$\frac{C_{pc}^{(n)}}{C_{pc}^{(n-1)}}=\frac{n}{\sum_{j=1}^{N}\eta_j\,(1-\pi_j^{(n-1)})}$$

e quindi

$$\pi_i^{(n)} = n\,\frac{\eta_i\,(1-\pi_i^{(n-1)})}{\sum_{j=1}^{N}\eta_j\,(1-\pi_j^{(n-1)})} \quad \textit{per ciascuna unità } i=1,\,\ldots,\,N. \tag{15.47}$$

Le N relazioni (15.47) forniscono uno schema ricorsivo numericamente efficiente per il calcolo delle $\pi_i^{(n)}$. Partendo infatti da

$$\begin{aligned}
&\pi_i^{(0)} = 0 \quad \textit{per ciascuna unità } i=1,\,\ldots,\,N;\\
&\pi_i^{(1)} = n\,\frac{\eta_i}{\sum_{j=1}^{N}\eta_j} \quad \textit{per ciascuna unità } i=1,\,\ldots,\,N
\end{aligned}$$

è facile calcolare successivamente le $\pi_i^{(2)}$, $\pi_i^{(3)}$, ... fino ad arrivare alla numerosità campionaria n.

In maniera simile si può costruire una relazione ricorsiva per il calcolo delle probabilità di inclusione del secondo ordine. Esattamente con la stesso approccio sopra utilizzato, non è difficile verificare (Esercizio 15.13) che

$$\pi_{ij}^{(n)} = \frac{\pi_i^{(n)}}{1-\pi_i^{(n-1)}}\,(\pi_j^{(n)}-\pi_{ij}^{(n-1)}) \tag{15.48}$$

da cui, partendo da espressioni ovvie per $\pi_{ij}^{(0)}$, $\pi_{ij}^{(1)}$, $\pi_{ij}^{(2)}$, e dopo aver calcolato le probabilità di inclusione del primo ordine, è immediato l'uso della (15.48) per il calcolo delle $\pi_{ij}^{(n)}$.

In linea di principio, i risultati ottenuti consentono anche di rispondere al quesito *Pa* 2. Per brevità non affronteremo questo problema, rinviando il lettore al volume di Tillé (2006).

15.5.2 Implementazione del disegno di Poisson condizionato

Per l'implementazione di un disegno di Poisson condizionato esistono diversi metodi, di cui esamineremo alcuni dei più importanti. Nel seguito considereremo un disegno di Poisson condizionato con probabilità dei campioni date dalla (15.39), in cui $\eta_i = \tau_i/(1-\tau_i)$ e $0 < \tau_i \leqslant 1$ per tutte le unità della popolazione.

Il metodo di implementazione più semplice consiste nello sfruttare il legame che il disegno di Poisson condizionato ha con il disegno di Poisson. L'idea di fondo è banale: si genera un campione in base ad un disegno di Poisson con $Pr(\delta(i;\ \mathbf{s}) = 1) = \tau_i$. Se il campione $\mathbf{s}$ generato ha numerosità n lo si accetta come campione di un disegno di Poisson condizionato, altrimenti lo si rifiuta e si ripete la procedura. Si tratta quindi di un algoritmo di accettazione condizionata, di seguito brevemente descritto.

- *Passo* 1. Porre $m = 0$; $\delta(1;\ \mathbf{s}) = \cdots = \delta(N;\ \mathbf{s}) = 0$; $i = 1$. Andare al Passo 2.
- *Passo* 2. Se $i > N$ andare al Passo 4. Se $m > n$ andare al Passo 1. Altrimenti andare al Passo 3.
- *Passo* 3. Generare un numero U con distribuzione uniforme in $[0, 1]$.
  Se $U \leqslant \tau_i$ porre $\delta(i;\ \mathbf{s}) = 1$. Incrementare m di 1. Incrementare i di 1. Andare al Passo 2.
- *Passo* 4. Se $m < n$ o se $m > n$ andare al Passo 1. Se $m = n$: stop. Il campione $\mathbf{s}$ è formato dalle n unità i tali che $\delta(i, \mathbf{s}) = 1$.

Per rendere il più conveniente possibile l'algoritmo dianzi esposto occorre che sia massima la probabilità di avere un campione di Poisson di numerosità n. Si può dimostrare (cfr. Tillé (2006)) che questo accade quando $\tau_i + \cdots + \tau_N = n$, ossia quando il disegno di Poisson condizionato è normalizzato.

L'algoritmo di selezione sopra esposto non è particolarmente efficiente sul piano computazionale. Un'utile alternativa, in genere computazionalmente assai vantaggiosa, consiste in un algoritmo di rigetto basato sulla generazione di un campione a partire dal disegno di Pareto. Per costruire l'algoritmo si osservi *in primis* che usando la notazione introdotta nella sezione precedente, per $\lambda_i = \tau_i$ e $\eta_i = \tau_i/(1-\tau_i)$, e se si indicano rispettivamente con $p_{pc}(\mathbf{s})$ e $p_{pa}(\mathbf{s})$ le probabilità dei campioni rispettivamente nel disegno di Poisson condizionato e in quello di Pareto, si ha

$$\begin{aligned} p_{pc}(\mathbf{s}) &= C_{pc} \left(\prod_{i=1}^{N} \eta_i^{\delta(i;\mathbf{s})} \right) \\ &= C'_{pc} \left(\prod_{i=1}^{N} \tau_i^{\delta(i;\mathbf{s})} (1-\tau_i)^{1-\delta(i;\mathbf{s})} \right) \\ &= C'_{pc}\, q(\mathbf{s}) \end{aligned}$$

$$
\begin{aligned}
p_{pa}(\mathbf{s}) &= C_{pa} \left(\prod_{i=1}^{N} \eta_i^{\delta(i;\,\mathbf{s})} \right) \left(\sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s}) \right) \\
&= \left(\prod_{i=1}^{N} \tau_i^{\delta(i;\,\mathbf{s})} \, (1-\tau_i)^{1-\delta(i;\,\mathbf{s})} \right) \left(\sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s}) \right) \\
&= q(\mathbf{s}) \left(\sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s}) \right)
\end{aligned}
$$

con C_{pc}, C'_{pc}, C_{pa} costanti opportune, g_k dato dalla (15.34) (e in via approssimata dalla (15.36)), e

$$
q(\mathbf{s}) = \prod_{i=1}^{N} \tau_i^{\delta(i;\,\mathbf{s})} \, (1-\tau_i)^{1-\delta(i;\,\mathbf{s})}.
$$

Posto quindi

$$
B \;=\; \frac{1}{\textit{Somma degli } n \textit{ più piccoli } g_k}
$$

si ha

$$
B \left(\sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s}) \right) \geqslant 1
$$

qualunque sia il campione $\mathbf{s}$, per cui dalle relazioni precedenti si ottiene

$$
\begin{aligned}
q(\mathbf{s}) &\leqslant B \, p_{pa}(\mathbf{s}) \\
\frac{q(\mathbf{s})}{B \, p_{pa}(\mathbf{s})} &= \frac{1}{B \sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s})}
\end{aligned}
$$

per ciascun campione $\mathbf{s}$. Utilizzando quanto detto al termine delle Sezione 12.8.3, si può generare un campione da un disegno di Poisson condizionato in base al seguente algoritmo di rigetto.

- *Passo* 1. `Inizializzazione. Generare un campione s da un disegno di Pareto con` $\lambda_1 = \tau_1$, ..., $\lambda_N = \tau_N$. `Andare al Passo 2.`
- *Passo* 2. `Generare una variabile aleatoria` U `con distribuzione uniforme in` $[0, 1]$. `Se`

$$
U \leqslant \frac{1}{B \sum_{k=1}^{N} g_k \, \delta(k;\,\mathbf{s})}
$$

`andare al Passo 3. Altrimenti, andare al Passo 1.`

- *Passo* 3. `Arresto. Accettare il campione s come selezionato dal disegno di Poisson condizionato.`

Le stesse idee possono anche essere usate per costruire uno schema di rigetto che implementa il disegno di Sampford. Si veda in proposito L'Esercizio 15.14.

15.6 Schemi di tipo scissorio*

Gli schemi di tipo scissorio (*splitting*) svolgono un ruolo non trascurabile nella costruzione di disegni campionari, sia sul piano teorico, in quanto si basano su un'idea semplice ed efficace, sia sul piano applicativo. Qui ci limiteremo solo a qualche cenno essenziale, rinviando il lettore curioso all'ottimo volume di Tillé (2006).

L'idea di base degli schemi scissori è semplice: *ad ogni passo si scinde il vettore delle N probabilità di inclusione in due o più parti vettori, fino ad arrivare ad un vettore composto solo da elementi pari a* 0 *oppure a* 1. *Gli elementi uguali a* 1 *corrispondono alle unità che fanno parte del campione, mentre quelli pari a* 0 *corrispondono alle unità che non fanno parte del campione.* Per ragioni di semplicità, si partirà da metodi che scindono il vettore delle probabilità di inclusione in due sole parti, per poi passare a metodi più generali.

*15.6.1 Schemi di scissione in due parti del vettore delle probabilità di inclusione**

Se $\pi_{01}, \ldots, \pi_{0N}$ sono le n probabilità di inclusione del primo ordine desiderate, con $n = \pi_{01} + \cdots + \pi_{0N}$ intero, definiamo il vettore iniziale

$$\boldsymbol{\pi}_0 = \begin{bmatrix} \pi_{01} \\ \pi_{02} \\ \cdots \\ \pi_{0N} \end{bmatrix}.$$

L'algoritmo di scissione in due vettori è di seguito esposto.

- *Passo* 0. Inizializzazione. Porre $t = 0$, $\boldsymbol{\pi}(0) = \boldsymbol{\pi}_0$ Andare al Passo 1.
- *Passo* 1. Se $\pi_i(t) = 0,\ 1$ per ciascuna unità $i = 1, \ldots, N$, andare al Passo 4. Altrimenti, andare al Passo 2.
- *Passo* 2. Scissione. Prendere due vettori a N componenti, $\boldsymbol{\pi}^1(t)$, $\boldsymbol{\pi}^2(t)$

$$\boldsymbol{\pi}^1(t) = \begin{bmatrix} \pi_1^1(t) \\ \pi_2^1(t) \\ \cdots \\ \pi_N^1(t) \end{bmatrix}, \quad \boldsymbol{\pi}^2(t) = \begin{bmatrix} \pi_1^2(t) \\ \pi_2^2(t) \\ \cdots \\ \pi_N^2(t) \end{bmatrix}$$

tali che

$$0 \leqslant \pi_i^1(t) \leqslant 1,\ \ 0 \leqslant \pi_i^2(t) \leqslant 1 \quad \textit{per ciascuna unità } i = 1, \ldots, N;$$
$$\textstyle\sum_{i=1}^N \pi_i^1(t) = \sum_{i=1}^N \pi_i^2(t) = n$$

e un numero reale $0 \leqslant \alpha(t) \leqslant 1$ tale che

$$\alpha(t)\, \boldsymbol{\pi}^1(t) + (1 - \alpha(t))\, \boldsymbol{\pi}^2(t) = \boldsymbol{\pi}(t).$$

Andare al Passo 3.

- *Passo* 3. Scelta casuale. Scegliere il vettore $\boldsymbol{\pi}^1(t)$ con probabilità $\alpha(t)$, e il vettore $\boldsymbol{\pi}^2(t)$ con probabilità $1 - \alpha(t)$. Porre il vettore scelto pari a $\boldsymbol{\pi}(t+1)$:

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}^1(t) & \textit{con probabilità } \alpha(t) \\ \boldsymbol{\pi}^2(t) & \textit{con probabilità } 1 - \alpha(t) \end{cases}.$$

  Incrementare t di 1. Andare al Passo 1.
- *Passo* 4. Arresto. Il campione s scelto è definito da $\delta(i; \mathbf{s}) = \pi_i(t)$ per ciascuna unità $i = 1, \ldots, N$.

La validità dei metodi di tipo scissorio poggia su due semplici considerazioni. In primo luogo, poiché si ha

$$\sum_{i=1}^{N} \pi_i(t) = n, \quad \textit{per qualunque } t = 1, 2, \ldots$$

è evidente che il campione $\mathbf{s}$ generato ha ampiezza n. In secondo luogo, da

$$\begin{aligned} E[\boldsymbol{\pi}(t+1) \,|\boldsymbol{\pi}(t),\ \boldsymbol{\pi}(t-1), \ldots, \boldsymbol{\pi}(0)] &= E[\boldsymbol{\pi}(t+1) \,|\boldsymbol{\pi}(t)] \\ &= \alpha(t)\, \boldsymbol{\pi}^1(t) + (1 - \alpha(t))\, \boldsymbol{\pi}^2(t) \\ &= \boldsymbol{\pi}(t) \end{aligned}$$

si desume che $E[\boldsymbol{\pi}(t)] = \boldsymbol{\pi}_0$. Detto quindi $\boldsymbol{\delta}(\mathbf{s})$ il vettore degli N indicatori $\delta(i; \mathbf{s})$ dianzi definiti, si ha

$$E[\boldsymbol{\delta}(\mathbf{s})] = \boldsymbol{\pi}_0$$

ovvero $\boldsymbol{\pi}_0$ è il vettore delle probabilità di inclusione del primo ordine.

In generale gli schemi di tipo scissorio sono molto semplici, e non difficili da programmare. Hanno però un difetto: a parte casi speciali, è molto difficile calcolare esattamente le probabilità di inclusione del secondo ordine. A peggiorare le cose si aggiunge il fatto spiacevole che alcune probabilità di inclusione del secondo ordine possono essere nulle; più in generale, è l'entropia dei disegni campionari corrispondenti a schemi scissori ad essere spesso bassa. Ciò rende poco o per nulla accurate le approssimazioni delle probabilità di inclusione del secondo ordine viste nella Sezione 12.7. Un rimedio (parziale) all'inconveniente della presenza di probabilità di inclusione del secondo ordine nulle consiste nel porre in ordine casuale le probabilità di inclusione $\pi_{01}, \ldots, \pi_{0N}$ nel vettore $\boldsymbol{\pi}_0$. In termini un po' più formali, si prende una permutazione "casuale" $(i_1, \ldots, i_N)$ di $(1, \ldots, N)$, in cui $i_1, \ldots, i_N$ sono N interi distinti, ciascuno dei quali può essere pari a 1, 2, ..., N. Si definisce il vettore $\boldsymbol{\pi}_0$ come

$$\boldsymbol{\pi}_0 = \begin{bmatrix} \pi_{0i_1} \\ \pi_{0i_2} \\ \cdots \\ \pi_{0i_N} \end{bmatrix}.$$

Si applica poi lo schema scissorio sopra esposto, fino ad arrivare al vettore "finale" (composto da n elementi uguali a 1 e $N-n$ elementi pari a 0)

$$\boldsymbol{\pi}(t) = \begin{bmatrix} \pi_{i_1}(t) \\ \pi_{i_2}(t) \\ \cdots \\ \pi_{i_N}(t) \end{bmatrix}.$$

Si riordinano infine gli elementi di $\boldsymbol{\pi}(t)$ in senso opposto rispetto a quanto fatto all'inizio, formando il campione $\mathbf{s}$ sulla base degli indicatori $\delta(i_1; \mathbf{s}) = \pi_{i_1}(t)$, $\delta(i_2; \mathbf{s}) = \pi_{i_2}(t)$, e così via. Questo modo di procedere fa in modo che le probabilità di inclusione del secondo ordine siano tutte positive, e in generale accresce l'entropia del disegno campionario. Tuttavia, non dice nulla sul calcolo effettivo delle probabilità di inclusione del secondo ordine.

Algoritmo del *pivot*

Una delle applicazioni dell'algoritmo di scissione è l'*algoritmo del pivot*. Esso si basa sullo scindere il vettore delle probabilità di inclusione in due vettori, modificando ad ogni passo solo due dei suoi elementi. In generale, all'iterazione t-ma indichiamo con i, j due unità della popolazione tali che $0 < \pi_i(t) < 1$ e $0 < \pi_j(t) < 1$. Se $\pi_i(t) + \pi_j(t) > 1$, poniamo

$$\pi_k^1(t) = \begin{cases} \pi_k(t) \;\; se\; k \neq i,\, j \\ 1 \;\; se\; k = i \\ \pi_i(t) + \pi_j(t) - 1 \;\; se\; k = j \end{cases}, \quad \pi_k^2(t) = \begin{cases} \pi_k(t) \;\; se\; k \neq i,\, j \\ 1 \;\; se\; k = j \\ \pi_i(t) + \pi_j(t) - 1 \;\; se\; k = i \end{cases} \tag{15.49}$$

$$\alpha(t) = \frac{1 - \pi_j(t)}{2 - \pi_i(t) - \pi_j(t)}.$$

Se invece $\pi_i(t) + \pi_j(t) < 1$, poniamo

$$\pi_k^1(t) = \begin{cases} \pi_k(t) \;\; se\; k \neq i,\, j \\ \pi_i(t) + \pi_j(t) \;\; se \;\; k = i \\ 0 \;\; se\; k = j \end{cases}, \quad \pi_k^2(t) = \begin{cases} \pi_k(t) \;\; se\; k \neq i,\, j \\ 0 \;\; se\; k = i \\ \pi_i(t) + \pi_j(t) \;\; se\; k = j \end{cases} \tag{15.50}$$

$$\alpha(t) = \frac{\pi_i(t)}{\pi_i(t) + \pi_j(t)}.$$

Nel primo caso viene scelto un vettore con un elemento uguale a 1, mentre nel secondo caso viene scelto un vettore con un elemento uguale a 0. È facile verificare (Esercizio 15.15) che in entrambi i casi si ha

$$\alpha(t)\, \boldsymbol{\pi}^1(t) + (1 - \alpha(t))\, \boldsymbol{\pi}^2(t) = \boldsymbol{\pi}(t). \tag{15.51}$$

Inoltre, è anche facile verificare che la somma delle componenti di $\boldsymbol{\pi}(t)$ è pari a n. Poiché ad ogni iterazione almeno un elemento $\pi_i(t)$ diventa pari a 0 o a 1, l'algoritmo termina al più in N iterazioni. Di seguito è fornita una descrizione dell'algoritmo per passi successivi.

- *Passo* 0. Inizializzazione. Porre $t = 0$, $\boldsymbol{\pi}(0) = \boldsymbol{\pi}_0$. Andare al Passo 1.
- *Passo* 1. Se $\pi_i(t) = 0,\ 1$ per ciascuna unità della popolazione, andare al Passo 6. Altrimenti, andare al Passo 2
- *Passo* 2.(Scelta di due unità) Prendere i più piccoli indici i, j tali che $0 < \pi_i(t) < 1$, $0 < \pi_j(t) < 1$. Se $\pi_i(t) + \pi_j(t) > 1$ andare al Passo 3. Altrimenti, andare al Passo 4.
- *Passo* 3. Definire $\boldsymbol{\pi}^1(t)$, $\boldsymbol{\pi}^2(t)$, $\alpha(t)$ come in (15.49). Andare al Passo 5.
- *Passo* 4. Definire $\boldsymbol{\pi}^1(t)$, $\boldsymbol{\pi}^2(t)$, $\alpha(t)$ come in (15.50). Andare al Passo 5.
- *Passo* 5. Scelta casuale. Scegliere il vettore $\boldsymbol{\pi}^1(t)$ con probabilità $\alpha(t)$, e il vettore $\boldsymbol{\pi}^2(t)$ con probabilità $1 - \alpha(t)$. Porre il vettore scelto pari a $\boldsymbol{\pi}(t+1)$. Incrementare t di 1. Andare al Passo 1.
- *Passo* 6. Arresto. Il campione $\mathbf{s}$ scelto è definito da $\delta(i; \mathbf{s}) = \pi_i(t)$ per ciascuna unità $i = 1, \ldots, N$.

Questa procedura ha il difetto di produrre alcune delle probabilità di inclusione del secondo ordine pari a 0. L'inconveniente può essere ovviato mediante una permutazione casuale delle π_{0i}, come sopra indicato. Molto più complicato è il calcolo delle probabilità di inclusione del secondo ordine, a meno di non ricorrere alle approssimazioni (in questo caso di dubbia qualità) della Sezione 12.7. Per ulteriori considerazioni sul tema si rinvia all'articolo di Deville e Tillé (1998).

Esempio 15.5. Consideriamo ancora la popolazione di $N = 10$ unità dell'Esempio 15.2, con $n = 5$ e le probabilità di inclusione del primo ordine desiderate sono $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$.

L'uso dell'algoritmo del pivot *senza permutazione iniziale delle unità* produce le probabilità di inclusione del secondo ordine riportate in Tabella 15.9.

Tabella 15.9 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – algoritmo del pivot

0.9000	0.6003	0.5498	0.4764	0.4458	0.4482	0.3595	0.3147	0.2249	0.1801
0.6003	0.7000	0.3503	0.3302	0.3378	0.3447	0.2785	0.2447	0.1748	0.1402
0.5498	0.3503	0.6500	0.2931	0.3102	0.3187	0.2581	0.2271	0.1623	0.1298
0.4764	0.3302	0.2931	0.5500	0.2062	0.2454	0.2111	0.1911	0.1366	0.1094
0.4458	0.3378	0.3102	0.2062	0.5000	0.1428	0.1680	0.1702	0.1216	0.0974
0.4482	0.3446	0.3187	0.2454	0.1428	0.5000	0.1249	0.1640	0.1172	0.0938
0.3595	0.2785	0.2581	0.2111	0.1680	0.1249	0.4000	0.0875	0.0625	0.0500
0.3147	0.2447	0.2271	0.1911	0.1702	0.1640	0.0875	0.3500	0.0000	0.0000
0.2249	0.1748	0.1623	0.1366	0.1216	0.1172	0.0625	0.0000	0.2500	0.0000
0.1801	0.1402	0.1298	0.1094	0.0974	0.0938	0.0500	0.0000	0.0000	0.2000

Tabella 15.10 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – algoritmo del pivot con permutazione casuale delle unità

0.9000	0.6212	0.5760	0.4848	0.4392	0.4392	0.3486	0.3043	0.2149	0.1703
0.6220	0.7000	0.4333	0.3595	0.3231	0.3231	0.2514	0.2169	0.1506	0.1201
0.5760	0.4333	0.6500	0.3291	0.2947	0.2947	0.2284	0.1963	0.1378	0.1095
0.4848	0.3595	0.3291	0.5500	0.2392	0.2392	0.1858	0.1610	0.1125	0.0891
0.4392	0.3231	0.2947	0.2392	0.5000	0.2125	0.1664	0.1441	0.1008	0.0798
0.4392	0.3231	0.2947	0.2392	0.2125	0.5000	0.1664	0.1441	0.1008	0.0798
0.3486	0.2514	0.2284	0.1858	0.1664	0.1664	0.4000	0.1126	0.0784	0.0618
0.3043	0.2169	0.1963	0.1610	0.1440	0.1440	0.1126	0.3502	0.0678	0.0534
0.2149	0.1506	0.1378	0.1125	0.1008	0.1008	0.0784	0.0678	0.2500	0.0364
0.1703	0.1201	0.1095	0.0891	0.0798	0.0798	0.0618	0.0534	0.0364	0.2000

Alcune delle π_{ij} sono uguali a zero, un fatto negativo per due ragioni. In primo luogo, non esiste uno stimatore corretto della varianza dello stimatore di Horvitz-Thompson. In secondo luogo, questo disegno ha un'entropia piccola. L'approssimazione π_{ij}^a delle π_{ij} sviluppata nella Sezione 12.7 produce, di conseguenza, risultati tutt'altro che buoni.

L'effettuare una permutazione casuale delle unità prima di applicare l'algoritmo del pivot, e il successivo "riordinamento inverso" dopo la selezione del campione, aumenta la "casualità" del disegno campionario, ossia la sua entropia. Le probabilità di inclusione del primo ordine delle unità restano ovviamente le stesse. Le probabilità di inclusione del secondo ordine sono riportate in Tabella 15.10.

Come è immediato verificare, le probabilità π_{ij} sono tutte positive. Inoltre, migliora di parecchio la qualità dell'approssimazione $\pi_{ij}^a \approx \pi_{ij}$. □

*15.6.2 Schemi di scissione in H parti del vettore delle probabilità di inclusione**

Una generalizzazione molto facile degli schemi in cui si scinde in due parti il vettore della probabilità di inclusione è quella in cui la scissione avviene in H parti, con H intero arbitrario. L'algoritmo di scissione in H vettori è esposto qui sotto.

- *Passo* 0. Inizializzazione. Porre $t = 0$, $\boldsymbol{\pi}(0) = \boldsymbol{\pi}_0$. Andare al Passo 1.
- *Passo* 1. Se $\pi_i(t) = 0$, 1 per ciascuna unità $i = 1, \ldots, N$, andare al Passo 4. Altrimenti, andare al Passo 2.
- *Passo* 2. Scissione. Prendere H vettori a N componenti, $\boldsymbol{\pi}^1(t)$, $\ldots$, $\boldsymbol{\pi}^H(t)$

$$\boldsymbol{\pi}^1(t) = \begin{bmatrix} \pi_1^1(t) \\ \pi_2^1(t) \\ \cdots \\ \pi_N^1(t) \end{bmatrix}, \ldots, \boldsymbol{\pi}^H(t) = \begin{bmatrix} \pi_1^H(t) \\ \pi_2^H(t) \\ \cdots \\ \pi_N^H(t) \end{bmatrix}$$

tali che

$$0 \leqslant \pi_i^1(t) \leqslant 1, \ldots, 0 \leqslant \pi_i^H(t) \leqslant 1 \quad \textit{per ciascuna unità } i = 1, \ldots, N;$$
$$\sum_{i=1}^N \pi_i^1(t) = \cdots = \sum_{i=1}^N \pi_i^H(t) = n$$

e H numeri reali $0 \leqslant \alpha_1(t) \leqslant 1, \ldots, 0 \leqslant \alpha_H(t) \leqslant 1$ tali che

$$\alpha_1(t) + \cdots + \alpha_H(t) = 1$$

e

$$\alpha_1(t)\, \boldsymbol{\pi}^1(t) + \cdots + \alpha_H(t)\, \boldsymbol{\pi}^H(t) = \boldsymbol{\pi}(t).$$

Andare al Passo 3.

– *Passo* 3. Scelta casuale. Scegliere il vettore $\boldsymbol{\pi}^1(t)$ con probabilità $\alpha_1(t)$, ..., il vettore $\boldsymbol{\pi}^H(t)$ con probabilità $\alpha_H(t)$. Porre il vettore scelto pari a $\boldsymbol{\pi}(t+1)$:

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}^1(t) & \textit{con probabilità } \alpha_1(t) \\ \cdots\cdots & \\ \boldsymbol{\pi}^H(t) & \textit{con probabilità } \alpha_H(t) \end{cases}.$$

Incrementare t di 1. Andare al Passo 1.

– *Passo* 4. Arresto. Il campione s scelto è definito da $\delta(i; \mathbf{s}) = \pi_i(t)$ per ciascuna unità $i = 1, \ldots, N$.

La validità dei metodi di scissione in H vettori si prova esattamente come nel caso di due vettori. Anche qui, i punti deboli sono la difficoltà di calcolare le probabilità di inclusione del secondo ordine e il fatto che in parecchi casi alcune di esse possono essere nulle.

Schema di Brewer

Una delle più importanti applicazioni dello schema generale di scissione è lo *schema di Brewer*, che si basa su una scissione del vettore delle probabilità di inclusione in N vettori. In generale, partendo da $\boldsymbol{\pi}(1) = \boldsymbol{\pi}_0$, il generico vettore $\boldsymbol{\pi}(t)$ ha $t-1$ elementi pari a 1, e i rimanenti $N-t+1$ proporzionali a π_{0i}. Per semplificare la sua costruzione è necessario definire preventivamente un vettore

$$\boldsymbol{d} = \begin{bmatrix} d_1 \\ d_2 \\ \cdots \\ d_N \end{bmatrix}$$

ciascuna componente del quale è pari o a 0, o a 1. Dato t, $t-1$ dei termini d_i sono uguali a 1 e $N-t+1$ sono uguali a 0. Formalmente, si ha:

$$\pi_i(t) = d_i + (1 - d_i)\, c_t\, \pi_{0i}, \quad \textit{per ciascuna unità } i = 1, \ldots, N \quad (15.52)$$

dove c_t è una costante di proporzionalità tale che la somma delle (15.52) sia pari a n. È facile verificare (Esercizio 15.16) che la costante c_t è pari a

$$c_t = \frac{n - t + 1}{n - \sum_{j=1}^{N} d_j \, \pi_{0j}}. \tag{15.53}$$

A partire da $\boldsymbol{\pi}(t)$ vengono costruiti gli N vettori $\boldsymbol{\pi}^1(t)$, …, $\boldsymbol{\pi}^N(t)$, in cui il generico $\boldsymbol{\pi}^k(t)$ ha pari a 1 le componenti in cui $d_i = 1$ e la k-ma, mentre tutte le altre sono proporzionali alle π_{0i}. Ciò significa che:

$$\pi_i^k(t) = \begin{cases} 1 \; se \; i = k \\ d_i + (1 - d_i) \, c_{t+1} \, \pi_{0i} \; se \; i \neq k \end{cases} \tag{15.54}$$

dove la costante di proporzionalità c_{t+1}, tale che la somma delle (15.54) sia pari a n, è eguale a

$$c_{t+1} = \frac{n - t}{n - \pi_{0k} - \sum_{j=1}^{N} d_j \, \pi_{0j}}. \tag{15.55}$$

La probabilità di scegliere il generico vettore $\boldsymbol{\pi}^k(t)$ è definita come

$$\alpha_k(t) = \frac{1}{S} \, (1 - d_k) \, \frac{\pi_{0k}(n - \pi_{0k} - \sum_{j=1}^{N} d_j \, \pi_{0j})}{n - \pi_{0k}(n - t + 1) - \sum_{j=1}^{N} d_j \, \pi_{0j}}, \quad k = 1, \ldots, N \tag{15.56}$$

con

$$S = \sum_{k=1}^{N} (1 - d_k) \, \frac{\pi_{0k}(n - \pi_{0k} - \sum_{j=1}^{N} d_j \, \pi_{0j})}{n - \pi_{0k}(n - t + 1) - \sum_{j=1}^{N} d_j \, \pi_{0j}}.$$

Si osservi che se $d_i = 1$ è $\alpha_i(t) = 0$. Poiché la somma delle componenti del vettore $\alpha_1(t)\boldsymbol{\pi}^1(t) + \cdots + \alpha_N(t)\boldsymbol{\pi}^N(t)$ è ovviamente pari a n, per verificare la validità del metodo basta provare che

$$\pi_i^1(t) \, \alpha_1(t) + \cdots + \pi_i^N(t) \, \alpha_N(t) = \pi_i(t) \tag{15.57}$$

per ciascuna unità della popolazione. Ora, se $d_i = 1$ la (15.57) è pari a 1. Se invece $d_i = 0$, la (15.57) risulta uguale (Esercizio 15.17) a

$$\begin{aligned} &\frac{1}{S} \left\{ \frac{\pi_{0i} \, (n - \pi_{0i} - \sum d_j \, \pi_{0j})}{n - (n - t + 1) \, \pi_{0i} - \sum d_j \, \pi_{0j}} \right. \\ &\left. + \sum_{\substack{k=1 \\ k \neq i}}^{N} (1 - d_k) \; \frac{\pi_{0k} \, (n - \pi_{0k} - \sum d_j \, \pi_{0j})}{n - (n - t + 1) \, \pi_{0k} - \sum d_j \, \pi_{0j}} \; \frac{(n - t) \, \pi_{0i}}{n - \pi_{0k} - \sum d_j \, \pi_{0j}} \right\} \\ &= c' \pi_{0i} \end{aligned} \tag{15.58}$$

con c' costante opportuna. Dalla (15.58) si desume che il vettore $\alpha_1(t)\boldsymbol{\pi}^1(t) + \cdots + \alpha_N(t)\boldsymbol{\pi}^N(t)$ ha esattamente la stessa struttura di $\boldsymbol{\pi}(t)$. Poiché la somma delle sue componenti è pari a n, se ne desume che $c' = c_t$, il che prova completamente la validità del metodo.

Lo schema di Brewer può essere posto in forma algoritmica nel modo seguente.

- *Passo* 0. Inizializzazione. Porre $t = 1$, $\boldsymbol{\pi}^1(t) = \cdots = \boldsymbol{\pi}^N(t) = \boldsymbol{\pi}_0$, $\alpha_1(t) = \cdots = \alpha_N(t) = 0$, $d_1 = \cdots = d_N = 0$. Andare al Passo 1.
- *Passo* 1. Se $t - 1 = n$ andare al Passo 5. Se $t - 1 < n$ andare al Passo 2.
- *Passo* 2. Scissione. Incrementare t di 1. Costruire gli N vettori $\boldsymbol{\pi}^1(t)$, ..., $\boldsymbol{\pi}^N(t)$. Andare al Passo 3.
- *Passo* 3. Probabilità di scelta. Calcolare le N probabilità (15.56). Andare al Passo 4.
- *Passo* 4. Scelta. Scegliere uno dei vettori $\boldsymbol{\pi}^1(t)$, ..., $\boldsymbol{\pi}^N(t)$ rispettivamente con probabilità $\alpha_1(t)$, ..., $\alpha_N(t)$. Se si sceglie il vettore $\boldsymbol{\pi}_k(t)$ porre $d_k = 1$. Andare al Passo 1.
- *Passo* 5. Arresto. Il campione $\mathbf{s}$ è formato dalle unità i tali che $d_i = 1$.

Lo schema di Brewer può anche essere visto come schema di n prove (iterazioni) successive. Ad ogni prova si seleziona, con probabilità (15.56), una delle unità della popolazione tali che $d_k = 0$; se i è l'unità scelta, si pone $d_i = 1$.

Per quanto riguarda le probabilità di inclusione del secondo ordine, il loro calcolo non è semplice. In Brewer (1975) viene fornita una formula ricorsiva, piuttosto pesante dal punto di vista computazionale. In alternativa, si possono usare le approssimazioni sviluppate nella Sezione 12.7.

Esempio 15.6. Nel caso di numerosità campionaria $n = 2$ lo schema di Brewer assume una forma molto semplice. I campioni sono combinazioni senza ripetizione di classe 2 delle unità della popolazione, e il generico $\mathbf{s} = \{i, j\}$ (tenendo conto che selezionare il campione $\{i, j\}$ significa che i è selezionata per prima e j per seconda o viceversa) ha probabilità

$$p(\mathbf{s}) = \frac{1}{\sum_{k=1}^{N} \frac{\pi_{0k}(2-\pi_{0k})}{2(1-\pi_{0k})}} \left\{ \frac{\pi_{0i}(2-\pi_{0i})}{2(1-\pi_{0i})} \frac{\pi_{0j}}{2-\pi_{0i}} + \frac{\pi_{0j}(2-\pi_{0j})}{2(1-\pi_{0j})} \frac{\pi_{0i}}{2-\pi_{0j}} \right\}$$
$$= \frac{1}{\sum_{k=1}^{N} (2-\pi_{0k}) \frac{\pi_{0k}}{1-\pi_{0k}}} \pi_{0i}\,\pi_{0j} \left(\frac{1}{1-\pi_{0i}} + \frac{1}{1-\pi_{0j}} \right).$$

Il confronto con i risultati dell'Esempio 15.1 mostra che per $n = 2$ il disegno di Brewer coincide con quello di Sampford. □

Esempio 15.7. Consideriamo di nuovo la popolazione di $N = 10$ unità dell'Esempio 15.2, con $n = 5$ e con probabilità di inclusione del primo ordine

Tabella 15.11 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – schema di Brewer

0.9000	0.6231	0.5770	0.4845	0.4392	0.4392	0.3489	0.3045	0.2156	0.1716
0.6231	0.7000	0.4305	0.3569	0.3220	0.3220	0.2525	0.2188	0.1533	0.1212
0.5770	0.4305	0.6500	0.3267	0.2942	0.2942	0.2299	0.1992	0.1395	0.1098
0.4845	0.3569	0.3267	0.5500	0.2408	0.2408	0.1862	0.1612	0.1129	0.0890
0.4390	0.3220	0.2942	0.2408	0.5000	0.2155	0.1665	0.1435	0.0994	0.0788
0.4390	0.3220	0.2942	0.2408	0.2155	0.5000	0.1665	0.1445	0.0994	0.0788
0.3489	0.2525	0.2299	0.1862	0.1665	0.1665	0.4000	0.1107	0.0771	0.0603
0.3045	0.2188	0.1992	0.1612	0.1440	0.1440	0.1107	0.3500	0.0663	0.0520
0.2156	0.1533	0.1395	0.1129	0.0996	0.0996	0.0772	0.0663	0.2500	0.0357
0.1716	0.1212	0.1097	0.0890	0.0787	0.0787	0.0603	0.0520	0.0357	0.2000

desiderate $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$.

L'uso dello schema di Brewer porta alle probabilità di inclusione del secondo ordine riportate in Tabella 15.11.

Dal confronto con quanto visto negli esempi precedenti, l'approssimazione $\pi_{ij}^a \approx \pi_{ij}$ fornisce buoni risultati. □

15.7 Schemi di tipo sistematico*

Gli schemi di tipo sistematico hanno un discreto interesse pratico, a causa della loro semplicità. Qui ci limiteremo al più semplice di essi proposto da Madow (1949); per approfondimenti si rinvia al volume di Tillé (2006), e ai relativi riferimenti bibliografici.

Indichiamo come al solito con $\pi_{01}, \ldots, \pi_{0N}$ le probabilità di inclusione del primo ordine desiderate, tali che $\pi_{01} + \cdots + \pi_{0N} = n$. Consideriamo inoltre le somme cumulate:

$$P_0 = 0,\ P_1 = \pi_{01},\ P_2 = \pi_{01} + \pi_{02},\ \ldots,\ P_N = \pi_{01} + \cdots + \pi_{0N} = n.$$

Lo schema sistematico di Madow è esposto qui di seguito in forma algoritmica.

- *Passo* 0. Inizializzazione. Calcolare le $N + 1$ quantità $P_0, P_1, \ldots P_N$. Andare al Passo 1.
- *Passo* 1. Generare una variabile aleatoria U con distribuzione uniforme in $[0, 1]$. Calcolare gli n numeri U, $U+1$, $U+2$, ..., $U+n-1$. Andare al Passo 2.
- *Passo* 2. Scelta. Includere nel campione s le unità i tali che

$$P_{i-1} < U + k \leqslant P_i \quad con\ k = 0, 1, \ldots, n-1.$$

Per verificare la validità dello schema dianzi illustrato, osserviamo che $k < U + k < k + 1$. Inoltre, essendo

$$0 = P_0 < P_1 < \cdots < P_{i-1} < P_i < P_{i+1} < \cdots < P_N = n$$

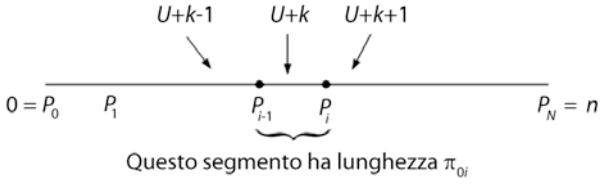


Fig. 15.1 Schema sistematico di Madow

e tenendo conto che ciascun intervallo $(P_{i-1}, P_i]$ ha lunghezza $0 < \pi_{0i} < 1$, ogni numero $U + k$ appartiene esattamente a uno e uno solo degli intervalli $(P_{i-1}, P_i]$. Ciò è anche illustrato in Fig. 15.1.

Se poi k e j sono due interi distinti (sempre compresi tra 0 e $n-1$) le due quantità $U+j$ e $U+k$ apparterranno a due differenti intervalli. Di conseguenza, i campioni selezionati con lo schema sistematico sopra descritto hanno tutti numerosità n.

Per calcolare le probabilità di inclusione del primo ordine iniziamo con l'osservare che se per P_{i-1} e P_i vi sono solo due possibilità, illustrate in Fig. 15.2:

A. per un qualche intero k si ha $k - 1 \leqslant P_{i-1} < P_i < k$;
B. per un qualche intero k si ha $P_{i-1} < k - 1 \leqslant P_i < k$.

Nel caso *A.* si ha (essendo $U + k - 1$ uniforme in $[k - 1, k]$)

$$\pi_i = Pr(P_{i-1} \leqslant U + k - 1 < P_i) = P_i - P_{i-1} = \pi_{0i}. \tag{15.59}$$

Nel caso *B.* si ha invece (essendo $U+k-2$ uniforme in $[k-2, k-1]$ e $U+k-1$ uniforme in $[k - 1, k]$)

$$\begin{aligned} \pi_i &= Pr(P_{i-1} \leqslant U + k - 2 < k - 1) + Pr(k - 1 \leqslant U + k - < P_i) \\ &= k - 1 - P_{i-1} + P_i - (k - 1) = \pi_{0i}. \end{aligned} \tag{15.60}$$

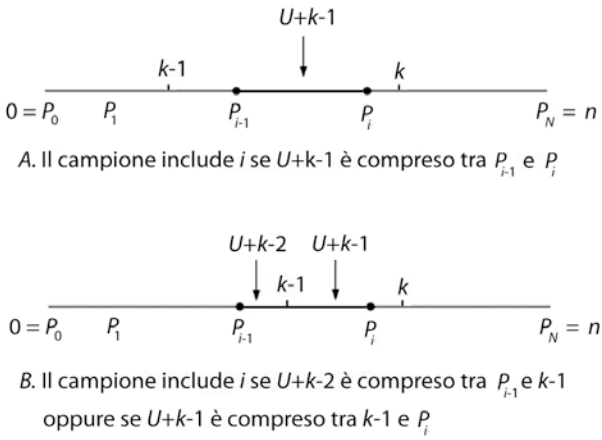


Fig. 15.2 Inclusione dell'unità i nel campione

Dalla (15.59) e (15.60) si conclude che la probabilità di inclusione dell'unità i è esattamente π_{0i}.

Le probabilità di inclusione del secondo ordine sono studiate in Connor (1966). Posto

$$\begin{aligned} P_{ij} &= \begin{cases} \sum_{k=i}^{j-1} \pi_{0k} & \textit{se } i < j \\ n - \sum_{k=j}^{i-1} \pi_{0k} & \textit{se } i > j \end{cases} \\ \lfloor P_{ij} \rfloor &= \textit{più grande intero} \leqslant P_{ij} \\ a_{ij} &= P_{ij} - \lfloor P_{ij} \rfloor \end{aligned}$$

si ha

$$\begin{aligned} \pi_{ij} = &\min\{\max(0, \pi_{0i} - a_{ij}),\ \pi_{0j}\} \\ &+ \min\{\pi_{0i},\ \max(0,\ a_{ij} + \pi_{0j} - 1)\}, \quad i < j. \end{aligned} \tag{15.61}$$

Il difetto principale dello schema sistematico appena esposto è che molte delle probabilità di inclusione del secondo ordine possono essere uguali a zero. All'inconveniente si può ovviare eseguendo una permutazione casuale delle unità della popolazione prima di applicare il metodo, e un "riordinamento inverso" delle unità stesse dopo la selezione del campione. Ovviamente, questo modo di procedere rende molto più complicato il calcolo delle π_{ij}. D'altra parte, esso migliora spesso la qualità dell'approssimazione π_{ij}^a sviluppata nella Sezione 16.6.

Esempio 15.8. Si consideri la popolazione di $N = 10$ unità dell'Esempio 15.2. Le probabilità di inclusione del primo ordine desiderate sono $\pi_{01} = 0.9$, $\pi_{02} = 0.7$, $\pi_{03} = 0.65$, $\pi_{04} = 0.55$, $\pi_{05} = 0.5$, $\pi_{06} = 0.5$, $\pi_{07} = 0.4$, $\pi_{08} = 0.35$, $\pi_{09} = 0.25$, $\pi_{010} = 0.2$, e la numerosità campionaria è $n = 5$.

L'uso dello schema sistematico di Madow, senza permutazione casuale delle unità della popolazione, porta alle probabilità di inclusione del secondo ordine riportate in Tabella 15.12.

Tabella 15.12 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – schema sistematico di Madow

0.9000	0.6000	0.5500	0.5500	0.4000	0.5000	0.3000	0.3500	0.2500	0.1000
0.6000	0.7000	0.3500	0.3500	0.4000	0.3000	0.3000	0.3500	0.0500	0.1000
0.5500	0.3500	0.6500	0.2000	0.4500	0.2000	0.4000	0.0500	0.2000	0.2000
0.5500	0.3500	0.2000	0.5500	0.0500	0.5000	0.0000	0.3000	0.2500	0.0000
0.4000	0.4000	0.4500	0.0500	0.5000	0.0000	0.4000	0.1000	0.0000	0.2000
0.5000	0.3000	0.2000	0.5000	0.0000	0.5000	0.0000	0.2500	0.2500	0.0000
0.3000	0.3000	0.4000	0.0000	0.4000	0.0000	0.4000	0.0000	0.0000	0.2000
0.3500	0.3500	0.0500	0.3000	0.1000	0.2500	0.0000	0.3500	0.0000	0.0000
0.2500	0.0500	0.2000	0.2500	0.0000	0.2500	0.0000	0.0000	0.2500	0.0000
0.1000	0.1000	0.2000	0.0000	0.2000	0.0000	0.2000	0.0000	0.0000	0.2000

Tabella 15.13 Valori esatti π_{ij} delle probabilità di inclusione del secondo ordine – schema sistematico di Madow con permutazione casuale delle unità

0.9000	0.6222	0.5775	0.4872	0.4411	0.4411	0.3490	0.3050	0.2126	0.1645
0.6222	0.7000	0.4413	0.3652	0.3273	0.3273	0.2485	0.2105	0.1408	0.1155
0.5775	0.4413	0.6500	0.3349	0.2975	0.2975	0.2229	0.1864	0.1337	0.1085
0.4872	0.3652	0.3349	0.5500	0.2342	0.2342	0.1808	0.1594	0.1134	0.0901
0.4411	0.3273	0.2975	0.2342	0.5000	0.2028	0.1654	0.1456	0.1039	0.0822
0.4411	0.3273	0.2975	0.2342	0.2028	0.5000	0.1654	0.1456	0.1039	0.0822
0.3490	0.2485	0.2229	0.1808	0.1654	0.1654	0.4000	0.1182	0.0842	0.0659
0.3050	0.2105	0.1864	0.1594	0.1456	0.1456	0.1182	0.3500	0.0728	0.0566
0.2126	0.1408	0.1337	0.1134	0.1039	0.1039	0.0842	0.0728	0.2500	0.0349
0.1645	0.1155	0.1085	0.0901	0.0822	0.0822	0.0659	0.0566	0.0349	0.2000

Alcune delle probabilità di inclusione del secondo ordine sono pari a zero, come già segnalato. Inoltre, l'approssimazione sviluppata nella Sezione 12.7 fornisce risultati piuttosto cattivi.

Nel caso in cui si effettui una permutazione casuale delle unità della popolazione, le probabilità di inclusione del secondo ordine sono quelle della Tabella 15.13. Si osservi come le π_{ij} siano tutte positive, e come l'approssimazione $\pi_{ij} \approx \pi_{ij}^{a}$ della Sezione 12.7 dia risultati soddisfacenti. □

15.8 Disegno campionario bilanciato*

*15.8.1 Definizione e aspetti di base**

Il disegno campionario di tipo bilanciato si basa su un'idea molto semplice: se sono disponibili *a priori* una o più variabili ausiliarie note sull'intera popolazione, il disegno campionario dovrebbe essere tale che le stime delle loro medie a livello campionario coincidano con le medie effettive a livello di popolazione. La logica di base è molto semplice: se il campione fornisce delle "buone" stime delle medie delle variabili ausiliarie, e se queste sono correlate con la variabile di interesse, anche la stima campionaria di quest'ultima dovrebbe essere "buona". Il primo tentativo di campionamento bilanciato è nel lavoro di Gini e Galvani (1929), dove, detto in termini moderni, si costruì un campione ragionato di 29 dei 214 distretti amministrativi in cui l'Italia era allora suddivisa. Un'analisi empirica basata sui dati del censimento del 1921, tuttavia, mostrò che i risultati delle stime ottenute su base campionaria erano buoni per le variabili di interesse molto correlate con quelle usate per costruire il campione, ma non buoni per altre variabili. Un esordio non certo incoraggiante! Il metodo fu fortemente criticato da Neyman (1934), ma a ben vedere le principali critiche di Neyman riguardavano il fatto che il campione fosse ragionato, ma non toccavano direttamente l'idea del "bilanciare" il disegno campionario sulla base di variabili ausiliarie. Detto in altri termini, anche se

all'epoca non era chiaro, il punto debole della proposta di Gini e Galvani stava nell'uso di un disegno campionario estremo di tipo ragionato, ma non nell'idea di bilanciamento *tout court*. Più di recente, Royall e Herson (1973) hanno fortemente sottolineato l'importanza del bilanciamento di un disegno campionario. Poiché all'epoca non esistevano metodi computazionalmente efficienti per costruire disegni bilanciati, essi proposero l'uso del disegno semplice senza ripetizione, che dovrebbe permettere una sorta di "bilanciamento in media". Il primo algoritmo realmente efficiente e di utilizzo generale è il *metodo del cubo*: Deville e Tillé (2004), Deville e Tillé (2005), Tillé (2006), Cap. 8.

La notazione che useremo in questa sezione è simile a quella introdotta per gli stimatori di tipo calibrazione. Si considerino p variabili ausiliarie $\mathcal{X}_1, \ldots, \mathcal{X}_p$, e sia x_{ik} il valore che la variabile $\mathcal{X}_k$ assume in corrispondenza dell'unità i ($i = 1, \ldots, N$; $k = 1, \ldots, p$).

Nel seguito si assumerà che i valori x_{ik} siano noti per tutte le unità della popolazione, per cui sono anche note le medie

$$\mu_{x_k} = \frac{1}{N} \sum_{i=1}^{N} x_{ik}; \quad k = 1, \ldots, p \tag{15.62}$$

delle variabili $\mathcal{X}_1, \ldots, \mathcal{X}_p$.

D'ora in avanti si assumerà *sempre* che il disegno campionario considerato sia non ordinato e senza ripetizioni. L'idea alla base del campionamento bilanciato, come dianzi accennato, è quella di utilizzare le variabili $\mathcal{X}_k$ per la selezione del campione. Formalmente, l'obiettivo è selezionare un campione in cui le stime delle medie delle variabili ausiliarie siano uguali alle medie effettive delle variabili stesse. Se (come sempre d'ora in poi) si fa riferimento allo stimatore di Horvitz-Thompson, il requisito che si richiede di soddisfare ad un disegno bilanciato è

$$\frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} x_{ik} = \mu_{x_k}, \quad k = 1, \ldots, p \tag{15.63}$$

per ciascun campione $\mathbf{s}$ dello spazio $\mathcal{S}$. Chiaramente, la (15.63) equivale a richiedere che il disegno campionario sia tale che gli stimatori di Horvitz-Thompson delle medie delle variabili ausiliarie abbiano varianza nulla:

$$V(t_{HT,\,x_k}) = 0, \quad k = 1, \ldots, p \tag{15.64}$$

dove si è posto

$$t_{HT,\,x_k} = \frac{1}{N} \sum_{i \in \mathbf{s}} \frac{1}{\pi_i} x_{ik}, \quad k = 1, \ldots, p. \tag{15.65}$$

Chiaramente l'idea sottostante al campionamento bilanciato è che se la condizione (15.63) è soddisfatta e le variabili ausiliarie $\mathcal{X}_k$ sono fortemente correlate con $\mathcal{Y}$, allora anche la stima di μ_y risulterà vicina alla "vera" media della popolazione μ_y.

Alla luce della definizione sopra data è evidente che il campionamento bilanciato può essere visto come una restrizione dello spazio campionario, in quanto soltanto i campioni che soddisfano la condizione (15.63) hanno probabilità di selezione positiva. Formalmente, lo spazio $\mathcal{S}$ dei campioni (di unità) è

$$\mathcal{S} = \{\mathbf{s} : \; t_{HT,\,x_k} = \mu_{x_k}, \;\; k = 1, \ldots, p\}\,.$$

Esempio 15.9. Se si usa un'unica variable ausiliaria $\mathcal{X}$ che per ciascuna unità i assume un valore pari alla probabilità di inclusione dell'unità stessa:

$$x_i = \pi_i \;\; \textit{per ciascuna unità } i = 1, \ldots N$$

allora si ha

$$\sum_{i\in\mathbf{s}} \frac{1}{\pi_i}\, x_i = n(\mathbf{s}) = \textit{ampiezza (effettiva) del campione } \mathbf{s}$$

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \pi_i = \overline{n} = \textit{ampiezza media (effettiva) campionaria}$$

per cui la (15.63) equivale a richiedere che

$$n(\mathbf{s}) = \overline{n} \;\; \textit{per ciascun campione } \mathbf{s}$$

ovvero che il disegno campionario sia ad ampiezza effettiva costante. □

Esempio 15.10. Si consideri un'unica variabile ausiliaria di bilanciamento che per tutte le unità della popolazione assume valore $x_i = 1$. La condizione di bilanciamento (15.63) si può scrivere come

$$\sum_{i\in\mathbf{s}} \frac{1}{\pi_i} = N \;\; \textit{per ciascun campione } \mathbf{s}. \tag{15.66}$$

Come già detto nel capitolo precedente, il peso da disegno $1/\pi_i$ si può interpretare come il numero di unità della popolazione "rappresentate" dall'unità campionaria i. Pertanto, si ha

$$\sum_{i\in\mathbf{s}} \frac{1}{\pi_i} = \begin{array}{l}\textit{numero di unità della popolazione rappresentate}\\ \textit{da quelle del campione } \mathbf{s}.\end{array}$$

La (15.66) equivale a richiedere che il numero di unità della popolazione rappresentate da quelle di ciascun campione sia uguale al numero effettivo di unità della popolazione, che è una condizione piuttosto intuitiva.

Questo esempio può anche essere interpretato in modo lievemente differente. La quantità

$$\sum_{i\in\mathbf{s}} \frac{1}{\pi_i}$$

è null'altro che lo stimatore di Horvitz-Thompson del numero totale N di unità della popolazione. La (15.66) richiede quindi che per ciascun campione esso coincida con il numero effettivo di unità della popolazione. □

Esempio 15.11. Anche la stratificazione può essere vista come una forma speciale di bilanciamento. Si supponga infatti che la popolazione di interesse sia suddivisa in M strati, rispettivamente di $N_1, N_2, \ldots, N_M$ unità. Per facilità di scrittura, in questo caso le unità della popolazione saranno etichettate con i numeri 1, 2, ..., N, e non con la solita doppia etichetta del tipo (*strato*, *etichetta di unità nello strato*). Definiamo poi M variabili $\mathcal{D}_1, \ldots, \mathcal{D}_M$, tali che per ciascuna unità della popolazione la variabile $\mathcal{D}_g$ assume il valore 1 se l'unità appartiene allo strato g, e il valore 0 in caso contrario. Si tratta, in sostanza, degli indicatori di appartenenza delle unità agli strati. In simboli:

$$d_{ig} = \begin{cases} 1 \textit{ se l'unità } i \textit{ appartiene allo strato } g \\ 0 \textit{ altrimenti} \end{cases}; \quad g = 1, \ldots, M;\ i = 1, \ldots, N.$$

Tenendo conto che $\sum_i d_{ig} = N_g$, le equazioni di bilanciamento assumono la forma:

$$\sum_{i \in \mathbf{s}} \frac{d_{ig}}{\pi_i} = N_g; \quad g = 1, \ldots, M \tag{15.67}$$

essendo $\sum_{i \in \mathbf{s}} d_{ig}/\pi_i$ lo stimatore di Horvitz-Thompson del numero di unità dello strato g. La (15.67) mostra quindi che le equazioni di bilanciamento implicano che lo stimatore di Horvitz-Thompson del numero di unità di ciascuno strato sia eguale al numero di unità dello strato stesso.

Detto $\mathbf{s}_g = \{i \in \mathbf{s} : \ d_{ig} = 1\}$ il sottocampione di $\mathbf{s}$ formato dalle unità dello strato g, le (15.67) si riscrivono come

$$\sum_{i \in \mathbf{s}_g} \frac{1}{\pi_i} = N_g; \quad g = 1, \ldots, M. \tag{15.68}$$

Nel caso speciale in cui le probabilità di inclusione delle unità di ciascuno strato g siano tutte uguali ad uno stesso numero, diciamo $\pi_{(g)}$, se si indica con $n(\mathbf{s}_g)$ il numero di unità campionarie dello strato g le (15.68) si riducono a

$$n(\mathbf{s}_g) = \pi_g N_g; \quad g = 1, \ldots, M$$

ossia il disegno campionario è sostanzialmente di tipo stratificato. □

In parecchi casi l'equazione di bilanciamento (15.63) può non essere esattamente soddisfatta. In tali circostanze è necessario procedere alla selezione di un campione approssimativamente bilanciato, ossia di un campione che soddisfi con buona approssimazione la (15.63).

Esempio 15.12. Supponiamo che $N = 6$, e che $\pi_i = 1/2$ per ciascuna unità della popolazione. Consideriamo due variabili ausiliarie di bilanciamento, $\mathcal{X}_1$, $\mathcal{X}_2$, definite nel modo seguente:

$$x_{i1} = \pi_i = \tfrac{1}{2}, \quad i = 1, \ldots, 6; \tag{15.69}$$

$$x_{1\,2} = x_{2\,2} = x_{3\,2} = 1, \; x_{4\,2} = x_{5\,2} = x_{6\,2} = 3. \tag{15.70}$$

La (15.69) implica che il disegno campionario debba avere ampiezza effettiva costante $n = 3$. Essendo poi le probabilità di inclusione del primo ordine tutte uguali, lo stimatore di Horvitz-Thompson si riduce alla media campionaria (di un campione di $n = 3$ unità). La variabile ausiliaria $\mathcal{X}_2$ ha media $\mu_{x_2} = 2$, ma, in conseguenza della (15.70), non può esserci nessun campione di 3 unità tale che la sua media campionaria sia pari a 2. Il problema è dovuto, in buona sostanza, al fatto che sia la numerosità della popolazione che quella del campione sono molto piccole. □

Esempio 15.13. Si consideri il *file* `spese_anziani.xls`, già visto nell'Esempio 10.5 del Capitolo 10, e in cui sono riportate diverse variabili relative a 250 comuni. La variabile di interesse, di cui si vuole stimare la media, è la spesa (media) per anziano sostenuta nell'anno 2011. Per ciascun comune sono note *a priori* la popolazione residente e la spesa per anziani sostenuta nell'anno 2009.

Per la selezione di un campione di $n = 20$ comuni consideriamo i disegni, tutti di tipo bilanciato, di seguito elencati.

BS1. Disegno bilanciato con ampiezza costante $n = 20$, probabilità di inclusione del primo ordine tutte uguali e variabile di bilanciamento spesa media per anziano sostenuta nell'anno 2009.

BS2. Disegno bilanciato con ampiezza costante $n = 20$, probabilità di inclusione del primo ordine proporzionali alla popolazione residente e variabile di bilanciamento spesa media per anziano sostenuta nell'anno 2009.

BS3. Disegno bilanciato con ampiezza costante $n = 20$, probabilità di inclusione del primo ordine proporzionali alla spesa media per anziano sostenuta nell'anno 2009, e variabile di bilanciamento popolazione residente.

BS4. Disegno bilanciato con ampiezza costante $n = 20$, probabilità di inclusione del primo ordine tutte uguali, e variabili di bilanciamento spesa media per anziano sostenuta nell'anno 2009 e popolazione residente.

BS5. Disegno con ampiezza costante $n = 20$, probabilità di inclusione del primo ordine proporzionali alla spesa media per anziano sostenuta nell'anno 2009.

In Tabella 15.14 sono riportate le varianze dello stimatore della spesa media per anziano nel 2011, per ciascuno dei cinque disegni campionari considerati.

Tabella 15.14 Deviazioni standard dello stimatore di Horvitz-Thompson per i disegno campionari BS1–BS5

Disegno campionario	*Varianze stimatore di Horvitz-Thompson*
BS1	72374.1
BS2	976993.4
BS3	14395.2
BS4	88145.9
BS5	13596.9

Il disegno con il comportamento peggiore è BS2, in cui le probabilità di inclusione del primo ordine dei comuni sono proporzionali alla corrispondente popolazione. Questo fatto si comprende bene se si fa riferimento alla Fig. 15.3, in cui in ascissa è riportata la popolazione dei comuni, e in ordinata il livello di spesa media per anziano nel 2011. Come subito si vede, i due caratteri sono molto lontani da una situazione di proporzionalità, così che lo scegliere probabilità di inclusione del primo ordine proporzionali al livello di popolazione determina una scarsissima efficienza dello stimatore di Horvitz-Thompson.

A rafforzare questa considerazione, è da rimarcare la debole correlazione tra i due caratteri spesa media per anziano nel 2011 e livello di popolazione, pari appena a -0.09.

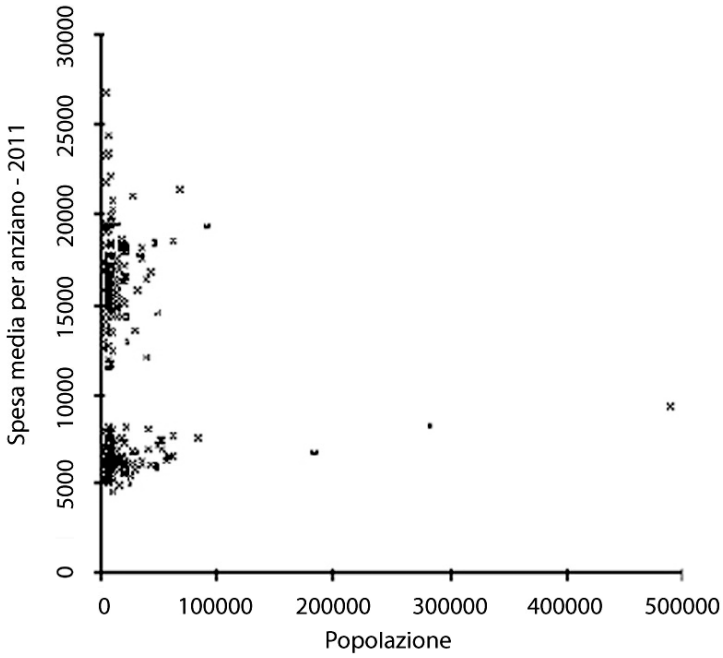


Fig. 15.3 Grafico della spesa media per anziano nel 2011 *vs* livello di popolazione

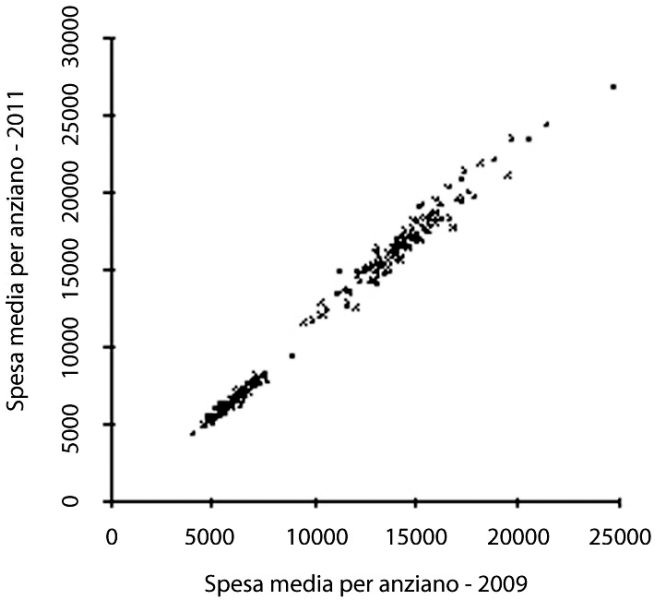


Fig. 15.4 Grafico della spesa media per anziano nel 2011 *vs* spesa media per anziano nel 2009

La scelta di usare probabilità di inclusione proporzionali alla spesa media per anziano nel 2009 fornisce invece risultati molto buoni, sia che non si usi la popolazione come variabile di bilanciamento (disegno BS3), sia che la si usi come variabile di bilanciamento (disegno BS5). Ciò si comprende bene se si considera la Fig. 15.4, in cui in ascissa è riportato il livello di spesa media per anziano nel 2009, e in ordinata quello nel 2011. Dalla figura appare chiaro come i due caratteri siano molto vicini ad una situazione di proporzionalità, per cui usare probabilità di inclusione del primo ordine proporzionali alla spesa media per anziano del 2009 si rivela una scelta molto efficace. Per la stessa ragione, tale variabile fornisce risultati migliori se usata per costruire le probabilità di inclusione piuttosto che come variabile di bilanciamento (con probabilità di inclusione costanti), come accade nei disegni BS1, BS4. □

*15.8.2 Il metodo del cubo**

Il *metodo del cubo* è il principale algoritmo per la selezione di un campione mediante un disegno bilanciato. La sua denominazione deriva dalla rappresentazione geometrica di un disegno campionario. Come visto nel Capitolo 12, un campione (non ordinato, senza ripetizioni) può essere rappresentato tramite il vettore delle N variabili indicatrici

$$\boldsymbol{\delta}(\mathbf{s}) = (\delta(1; \mathbf{s}), \delta(2; \mathbf{s}), \ldots, \delta(N; \mathbf{s})). \tag{15.71}$$

Geometricamente ogni vettore $\boldsymbol{\delta}(\mathbf{s})$ può essere visto come uno dei 2^N vertici dell'ipercubo N-dimensionale $[0, 1]^N$. Ogni vettore $\boldsymbol{\delta}(\mathbf{s})$ identifica un vertice dell'ipercubo $[0, 1]^N$, e di conseguenza il corrispondente campione $\mathbf{s}$.

L'interpretazione geometrica dei disegni campionari è già stata studiata nel Capitolo 12. Un disegno campionario caratterizzato da probabilità di inclusione π_i, $i = 1, \ldots, N$, assegna una probabilità $p(\boldsymbol{\delta}(\mathbf{s})) = p(\mathbf{s})$ di selezione a ogni vertice dell'ipercubo, in maniera tale che

$$\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s})\,\boldsymbol{\delta}(\mathbf{s}) = \boldsymbol{\pi} \tag{15.72}$$

essendo $\boldsymbol{\pi}$ il vettore delle probabilità di inclusione. Geometricamente un disegno campionario porta ad esprimere il vettore $\boldsymbol{\pi}$ come una combinazione lineare convessa dei vertici (di probabilità positiva) dell'ipercubo $[0, 1]^N$. Di conseguenza, selezionare un campione significa scegliere un vertice dell'ipercubo a partire da un vettore $\boldsymbol{\pi}$, in modo tale da soddisfare l'equazione di bilanciamento (15.63).

Per capire un po' meglio il significato geometrico delle (15.63), definiamo la matrice $p \times N$ (in cui ogni riga si riferisce ad una variabile e ogni colonna ad un'unità):

$$\mathbf{A} = \begin{bmatrix} \frac{x_{11}}{\pi_1} & \frac{x_{21}}{\pi_2} & \ldots & \frac{x_{N1}}{\pi_N} \\ \frac{x_{12}}{\pi_1} & \frac{x_{22}}{\pi_2} & \ldots & \frac{x_{N2}}{\pi_N} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{x_{1p}}{\pi_1} & \frac{x_{2p}}{\pi_2} & \ldots & \frac{x_{Np}}{\pi_N} \end{bmatrix}. \tag{15.73}$$

Le p equazioni (15.63), come facilmente si vede, possono essere scritte come

$$\mathbf{A}\,\boldsymbol{\delta}(\mathbf{s}) = \mathbf{A}\,\boldsymbol{\pi}$$

ossia come

$$\mathbf{A}\,(\boldsymbol{\delta}(\mathbf{s}) - \boldsymbol{\pi}) = \mathbf{0}_p \tag{15.74}$$

essendo $\mathbf{0}_p$ un vettore di p numeri 0. Quindi, geometricamente le equazioni di bilanciamento stabiliscono che il vettore $\boldsymbol{\delta}(\mathbf{s})$ deve appartenere ad un sottospazio lineare (affine) di $\mathbb{R}^N$ di dimensione $N - p$. Precisamente, detto

$$\mathrm{Ker}\mathbf{A} = \{\boldsymbol{u} \in \mathbb{R}^N : \ \mathbf{A}\,\boldsymbol{u} = \mathbf{0}_p\}$$

il *nucleo* (*kernel*) della matrice $\mathbf{A}$, il vettore $\boldsymbol{\delta}(\mathbf{s})$ deve essere tale che $\boldsymbol{\delta}(\mathbf{s}) - \boldsymbol{\pi}$ appartiene a Ker$\mathbf{A}$.

Il metodo del cubo si compone di due fasi: la *fase di volo* e la *fase di atterraggio*. Nella fase di volo si sceglie, in maniera "casuale", un vettore $\boldsymbol{\delta}(\mathbf{s})$ tale che $\boldsymbol{\delta}(\mathbf{s}) - \boldsymbol{\pi}$ appartenga a Ker$\mathbf{A}$. Tale vettore non ha necessariamente componenti tutte uguali a 0 o a 1. Nella fase di atterraggio si rilassano i vincoli delle equazioni di bilanciamento in modo da ottenere un vettore a componenti

0 o 1. Un'ultima notazione prima di esporre il metodo del cubo. Si tratta, in sostanza, di uno schema che implementa un disegno in cui le probabilità di inclusione del primo ordine sono prefissate, e vanno anche soddisfatti i vincoli di bilanciamento. Con tale metodo il calcolo delle probabilità di inclusione del secondo ordine è un problema numericamente quasi insormontabile. Buoni risultati, comunque, sono offerti sia dalle approssimazioni sviluppate nel Capitolo 12, sia dalle corrispondenti approssimazioni della varianza dello stimatore di Horvitz-Thompson.

Fase di volo

La fase di volo, come anticipato, genera un vettore $\boldsymbol{\delta}(\mathbf{s})$ che soddisfi sia le equazioni di bilanciamento (15.74) che la (15.72) (necessaria per rispettare il vincolo sulle probabilità di inclusione del primo ordine). Qui di seguito è brevemente presentato l'algoritmo principale per l'implementazione della fase di volo.

- *Passo* 0. Inizializzazione. Porre $t = 0$, $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$. Andare al Passo 1.
- *Passo* 1. Generare (casualmente o meno) un vettore $\boldsymbol{u}(t) \neq \mathbf{0}_p$ appartenente a $\mathrm{Ker}\mathbf{A}$ e avente componenti $u_1(t), \ldots, u_N(t)$ tali che $u_i(t) = 0$ se $\pi_i(t)$ è intero. Se $\boldsymbol{u}(t)$ ha componenti tutte uguali a 0 andare al passo 4. Altrimenti, andare al Passo 2.
- *Passo* 2. Calcolare i numeri $\lambda_1^*(t)$, $\lambda_2^*(t)$ come:

$$\lambda_1^*(t) = \textit{più grande valore di } \lambda_1(t) \textit{ tale che } \mathbf{0}_p \leqslant \boldsymbol{\pi}(t) + \lambda_1(t)\,\boldsymbol{u}(t) \leqslant \mathbf{1}_p$$
$$\lambda_2^*(t) = \textit{più grande valore di } \lambda_2(t) \textit{ tale che } \mathbf{0}_p \leqslant \boldsymbol{\pi}(t) - \lambda_2(t)\,\boldsymbol{u}(t) \leqslant \mathbf{1}_p$$

  essendo $\mathbf{1}_p$ un vettore di p componenti tutte uguali a 1, ed essendo le disuguaglianze tra vettori intese come valide componente per componente. Si osservi che $\lambda_1^*(t) > 0$, $\lambda_2^*(t) > 0$. Andare al Passo 3.
- *Passo* 3. Scegliere

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\,\boldsymbol{u}(t) \textit{ con probabilità } q(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\,\boldsymbol{u}(t) \textit{ con probabilità } 1 - q(t) \end{cases}$$

  con $q(t) = \frac{\lambda_2^*(t)}{\lambda_1^*(t)+\lambda_2^*(t)}$. Incrementare t di 1 e andare al Passo 1.
- *Passo* 4. Arresto. Stop: porre $\widetilde{\boldsymbol{\delta}} = \boldsymbol{\pi}(t)$.

Come mostrato in Deville e Tillé (2004), il vettore $\widetilde{\boldsymbol{\delta}}$ ha valore atteso pari al vettore $\boldsymbol{\pi}$ delle probabilità di inclusione del primo ordine, ed inoltre soddisfa le equazioni di bilanciamento (15.74). Le componenti del vettore $\widetilde{\boldsymbol{\delta}}$ sono tutte comprese tra 0 e 1 ed al più p di esse non sono uguali né a 0, né a 1. Se questo accade, è necessario passare alla fase di atterraggio. Altrimenti, basta porre $\boldsymbol{\delta}(\mathbf{s}) = \widetilde{\boldsymbol{\delta}}$ e la generazione del campione $\mathbf{s}$ è completa.

Prima di passere a discutere brevemente la fase di atterraggio, un'ultima notazione. Il Passo 1 è molto oneroso sul piano computazionale. Un algoritmo computazionalmente per la sua realizzazione è nel lavoro di Chauvet e Tillé (2006).

Fase di atterraggio

La fase di atterraggio si rende necessaria quando il vettore $\widetilde{\boldsymbol{\delta}}$ ottenuto nella fase di volo non ha tutte le componenti uguali a 0 o a 1, ma contiene qualche elemento $0 < \widetilde{\delta}_i < 1$. L'idea di fondo della fase di atterraggio è quella di rilassare i vincoli di bilanciamento, in modo da ottenere nella fase di volo un vettore $\widetilde{\boldsymbol{\delta}}$ a componenti pari a 0 o 1. Di seguito sono elencati i più semplici metodi di atterraggio.

- *Eliminazione di vincoli.* Questo metodo consiste nel dare un ordine di importanza ai vincoli, e nell'eliminare gli stessi partendo da quello meno importante. Ogni volta che si elimina un vincolo, si rieffettua la fase di volo, fino a quando non si giunge ad un vettore $\widehat{\boldsymbol{\delta}}$ a componenti tutte pari a 0 o a 1.
- *Minimizzazione di funzione di costo - 1.* Dato il vettore $\widetilde{\boldsymbol{\delta}}$ ottenuto al termine della fase di volo, si considerano le sue componenti non uguali né a 0 e né a 1, che in generale saranno in numero di $q \leqslant p$. Si costruiscono poi tutti i vettori (in totale sono 2^q costruiti a partire da $\widetilde{\boldsymbol{\delta}}$ e ponendo uguali a 0 o a 1 tutte le sue componenti che non soddisfano tale requisito. Formalmente, si costruiscono i vettori $\widetilde{\boldsymbol{\delta}}^j$, $j = 1, \ldots, 2^q$ tali che $\widetilde{\delta}_i^j = \widetilde{\delta}_i$ se $\widetilde{\delta}_i$ è pari a 0 o a 1, e $\widetilde{\delta}_i^j \in \{0, 1\}$ se $0 < \widetilde{\delta}_i < 1$, con $i = 1, \ldots, N$. Ciascuno dei vettori $\widetilde{\boldsymbol{\delta}}^j$ corrisponde ad un campione $\mathbf{s}_j$, $j = 1, \ldots, 2^q$. Sia $t^j_{HT,\,x_k}$ il valore dello stimatore di Horvitz-Thompson di μ_{x_k} calcolato in corrispondenza del campione $\mathbf{s}_j$. Il *costo* di $\mathbf{s}_j$ è definito come

$$C(\mathbf{s}_j) = \sum_{k=1}^{p} \left| \frac{t^j_{HT,\,x_k} - \mu_{x_k}}{\mu_{x_k}} \right| .$$

  L'idea di base è di scegliere "casualmente" uno tra $\mathbf{s}_1, \ldots, \mathbf{s}_{2^q}$, in modo che il generico $\mathbf{s}_j$ abbia probabilità $\widetilde{p}(\mathbf{s}_j)$ di essere selezionato. Le probabilità $\widetilde{p}(\mathbf{s}_j)$ sono determinate in modo da minimizzare il costo atteso

$$\sum_j C(\mathbf{s}_j)\, \widetilde{p}(\mathbf{s}_j)$$

  e in modo da soddisfare i vincoli

$$\sum_j \widetilde{p}(\mathbf{s}_j) = 1$$

$$\sum_j \widetilde{\delta}_i^j\, \widetilde{p}(\mathbf{s}_j) = \widetilde{\delta}_i; \quad i = 1, \ldots, N.$$

– *Minimizzazione di funzione di costo - 2.* Si tratta di un metodo simile al precedente, ma si richiede anche che il campione selezionato abbia ampiezza (effettiva) pari a $n = \pi_1 + \cdots + \pi_N$.

Per approfondimenti sul metodo del cubo si rinvia a Deville e Tillé (2004), Tillé (2006), Cap. 8.

15.9 L'utilizzo di R nel campionamento da popolazioni finite

R è un software statistico scaricabile gratuitamente dalla R home-page, *http://www.r-project.org*. R è programma Open Source, ognuno può avere accesso al suo codice interno ed, eventualmente, proporne modifiche. La caratteristica principale di R è la sua modularità: tutte le sue funzioni sono contenute in dei pacchetti ad ognuno dei quali è dedicato un compito specifico. La distribuzione di base di R include un certo numero di pacchetti che sono necessari per un funzionamento minimale del sistema, ogni altro pacchetto non presente nella versione di R può essere installato digitando nella console di R il comando

install.packages("nome del pacchetto").

Nell'ambito della teoria del campionamento da popolazioni finite Tillé e Matei (2009) hanno sviluppato il pacchetto *sampling* in cui sono implementati diversi disegni di campionamento e diversi metodi di stima. Se si vuole installare il pacchetto *sampling* durante una sessione di R bisogna digitare nella console di R il comando

install.packages("sampling").

Una volta installato il pacchetto è necessario richiamare il pacchetto all'interno della sessione di lavoro attraverso il comando

library(sampling).

Se si vogliono reperire informazioni specifiche sul contenuto del pacchetto *sampling* basterà digitare

library(help=sampling)

o in alternativa

help(package=sampling).

Tali comandi causano l'apertura di una finestra che riporta tutte le funzioni e i dataset specifici del pacchetto. Per ottenere informazioni su una funzione del pacchetto basterà digitare nel prompt il comando

help(nome funzione).

Ad esempio se si vogliono informazioni sulla funzione *srswor* che consente la selezione di un campione secondo un disegno semplice senza ripetizione basterà digitare nel prompt

help(srswor).

Un campione casuale semplice può essere selezionato usando la funzione *srswor* per un campionamento senza ripetizione e *srswr* per un campionamento con ripetizione. Per quanto riguarda i disegni di campionamento a probabilità variabili alcuni dei disegni implementati nel pacchetto sono elencati di seguito:

- disegno di Brewer, implementato nella funzione *UPbrewer*;
- disegno di massima entropia, implementato nella funzione *UPmaxentropy*;
- disegno di Midzuno-Lahiri, implementato nella funzione *UPmidzuno*;
- metodo del pivot, implementato nella funzione *UPpivotal*;
- disegno di tipo Pareto, implementato nella funzione *UPopips*;
- disegno di Poisson, implementato nella funzione *UPpoisson*;
- disegno sistematico, implementato nella funzione *UPsystematic*;
- disegno di Sampford, implementato nella funzione *UPsampford*.

È inoltre possibile selezionare un campione utilizzando:

- un disegno stratificato, con probabilità di inclusione uguali o diverse;
- un disegno a grappolo, con probabilità di inclusione uguali o diverse;
- un disegno a due o più stadi di campionamento, con probabilità di inclusione uguali o diverse in ogni stadio.

Nel pacchetto *sampling* sono anche implementate funzioni che consentono il calcolo delle probabilità di inclusione del secondo ordine per il disegno di massima entropia, il disegno di Midzuno, il disegno sistematico.

Il pacchetto *sampling* consente inoltre la selezione di un campione bilanciato attraverso il metodo del cubo; l'algoritmo è implementato nella funzione *samplecube*.

Per quanto riguarda i metodi di stima, le funzioni *HTestimator* e *HTstrata* consentono di stimare, rispettivamente con disegno semplice e con disegno stratificato, il totale della popolazione utilizzando lo stimatore di Horvitz-Thompson. Con riferimento a stimatori che si basano sull'utilizzo di informazione ausiliarie lo stimatore per quoziente è implementato nelle funzioni *ratioest*,*ratioest_strata*, e lo stimatore di regressione è implementato nelle funzioni *regest*, *regest_strata* (rispettivamente per il disegno ssr e per quello stratificato). Il pacchetto consente inoltre sia il calcolo dei pesi per la definizione dello stimatore calibrato attraverso la funzione *calib* sia il calcolo dello stimatore calibrato e della sua varianza attraverso l'utilizzo della funzione *calest*. La post stratificazione è implementata nella funzione *poststrata*, e lo stimatore post stratificato è calcolato usando la funzione *postest*.

Esercizi

15.1. Siano $(\mathcal{S}_1,\ p_1(\cdot))$ $(\mathcal{S}_2,\ p_2(\cdot))$ due disegni campionari ad ampiezza effettiva costante n, e si indichino con $\pi_i^{(1)}$, $\pi_i^{(2)}$, $\pi_{ij}^{(1)}$ $\pi_{ij}^{(2)}$ le loro probabilità di inclusione, rispettivamente del primo e del secondo ordine. Si indichino inoltre con $V_1(t_{HT})$, $V_2(t_{HT})$ le corrispondenti varianze dello stimatore di Horvitz-Thompson di μ_y.

a. Provare che se $\pi_1^{(1)} = \pi_1^{(2)}, \ldots, \pi_N^{(1)} = \pi_N^{(2)}$, e se $V_1(t_{HT}) \leqslant V_2(t_{HT})$ per qualche vettore $\boldsymbol{Y}_N$, allora deve essere $V_1(t_{HT}) \geqslant V_2(t_{HT})$ per qualche altro vettore $\boldsymbol{Y}_N$.

Suggerimento. Detta Ψ_1 la matrice $N \times N$ di elementi $\pi_{ij}^{(1)}/(\pi_i^{(1)}\pi_j^{(1)}) - 1$, e detta Ψ_2 la matrice $N \times N$ di elementi $\pi_{ij}^{(2)}/(\pi_i^{(2)}\pi_j^{(2)}) - 1$, si ha $V_1(t_{HT}) = Y'_N \Psi_1 Y_N/N^2$, e $V_2(t_{HT}) = Y'_N \Psi_2 Y_N/N^2$. Se fosse $V_1(t_{HT}) \leqslant V_2(t_{HT})$ per ogni vettore $Y_N \in \mathbb{R}^N$, si avrebbe $Y'_N(\Psi_2 - \Psi_1)Y_N \geqslant 0$ per ogni $Y_N \in \mathbb{R}^N$, così che la matrice $\Psi_2 - \Psi_1$ sarebbe semidefinita positiva. I suoi autovalori, quindi, sarebbero tutti non negativi. D'altra parte, se $\pi_i^{(1)} = \pi_i^{(2)}$, $i = 1, \ldots, N$, la matrice $\Psi_2 - \Psi_1$ ha diagonale principale composta da N zeri, e quindi anche i suoi autovalori devono essere uguali a zero (perché?). Ma allora anche gli altri elementi di $\Psi_2 - \Psi_1$ dovrebbero essere tutti nulli.

b. Dedurre da *a.* che non esiste nessun disegno ad ampiezza effettiva costante e con prefissate probabilità di inclusione del primo ordine che rende minima la varianza dello stimatore di Horvitz-Thompson della media della popolazione.

15.2. Provare che tra tutti i disegni campionari non ordinati, senza ripetizioni, e con prefissate probabilità di inclusione del primo ordine $\pi_{01}, \ldots, \pi_{0N}$, il disegno di Poisson è quello di entropia massima.

Suggerimento. Bisogna minimizzare la quantità $\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \log p(\mathbf{s})$, subordinatamente ai vincoli $\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) = 1$, $\sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\mathbf{s})\, p(\mathbf{s}) = \pi_{0i}$, $i = 1, \ldots, N$. La funzione Lagrangiana assume la forma: $\mathcal{L} = \sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \log p(\mathbf{s}) - \sum_{i=1}^N \lambda_i(\sum_{\mathbf{s}\in\mathcal{S}} \delta(i;\mathbf{s})\, p(\mathbf{s}) - \pi_{0i}) - \lambda_{N+1}(\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) - 1)$. Derivando rispetto a $p(\mathbf{s})$ e annullando le derivate, si ha poi $\partial\mathcal{L}/\partial p(\mathbf{s}) = \log p(\mathbf{s}) + 1 - \sum_{i=1}^N \lambda_i \delta(i;\mathbf{s}) - \lambda_{N+1} = 0$, da cui $\log p(\mathbf{s}) = cost + \sum_{i=1}^N \lambda_i \delta(i;\mathbf{s})$, e quindi $p(\mathbf{s}) = p(\delta(1;\mathbf{s}), \ldots, \delta(N;\mathbf{s})) = C \prod_{i=1}^N \theta_i^{\delta(1;\mathbf{s})}$, essendo $\theta_i = \exp(\lambda_i)$, e C una costante opportuna. Quest'ultima relazione mostra che le variabili aleatorie $\delta(1;\mathbf{s})$ sono indipendenti, e quindi il disegno deve essere di Poisson. Poiché deve anche essere $Pr(\delta(i;\mathbf{s}) = \pi_{0i}/(1-\pi_{0i})$, si conclude che $\theta_i = \pi_{0i}$, e $C = C_{po}$.

15.3. Verificare che per il disegno campionario di Poisson lo stimatore di Horvitz-Thompson t_{HT} ha varianza (15.13).

15.4. (Stimatore di Horvitz-Thompson con disegno di Poisson) Dato un disegno di Poisson con probabilità di inclusione del primo ordine $\pi_{01} \ldots, \pi_{0N}$, si consideri lo stimatore t_{HT} di Horvitz-Thompson di μ_y.

a. Verificare che il rapporto $E[t_{HT}^2]/\mu_y^2 = 1 + V(t_{HT})/\mu_y^2$, qualunque sia μ_y fissato, raggiunge il suo valore minimo per $y_i = K\pi_{0i}/(1-\pi_{0i})$, $i = 1, \ldots, N$, con $K = \mu_y / \left(\frac{1}{N}\sum_{i=1}^N \frac{\pi_{0i}}{1-\pi_{0i}}\right)$.

b. Verificare che il valore ottimo della costante di contrazione è

$$c^* = \sum_{i=1}^N \frac{\pi_{0i}}{1-\pi_{0i}} \Bigg/ \left(1 + \sum_{i=1}^N \frac{\pi_{0i}}{1-\pi_{0i}}\right).$$

15.5. (Costante di normalizzazione per il disegno di Sampford) Con riferimento al disegno di Sampford, si definiscano le seguenti grandezze:

a. $\mathcal{C}_{N-1,m}(\overline{i})$: classe di tutte le combinazioni senza ripetizioni di m unità di tutta la popolazione I_N privata di i, $I_N \setminus \{i\}$.
b. $g(n, r, i) = \sum_{\mathbf{c}\in\mathcal{C}_{N,n-r}(\overline{i})} \left\{\prod_{j=1}^N \omega_j^{\delta(j;\mathbf{c})}\right\} \left(n - rp_i - \sum_{k=1}^N p_k \delta(k;\mathbf{c})\right)$.
c. $h(n, r, i) = \sum_{\mathbf{c}\in\mathcal{C}_{N,n-r}(\overline{i})} \left\{\prod_{j=1}^N \omega_j^{\delta(j;\mathbf{c})}\right\} \left(\sum_{k=1}^N (1-p_k) \delta(k;\mathbf{c})\right)$.
d. $D_m(\overline{i}) = \sum_{\mathbf{c}\in\mathcal{C}_{N-1,m}(\overline{i})} \left\{\prod_{j=1}^N \omega_j^{\delta(j;\mathbf{c})}\right\}$.

Provare le seguenti relazioni.

a. $g(n, r, i) = r(1-p_i) D_{nr}(\overline{i}) + h(n, r, i)$.
b. $h(n, r, i) = \sum_{\mathbf{c}\in\mathcal{C}_{N-1,n-r}(\overline{i})} \sum_{k\in\mathbf{c}} p_k \left(\prod_{j\in\mathbf{c};\, j\neq k} \omega_j\right)$.
c. $h(n, r, i) = g(n, r+1, i) + r\pi_{0i} D_{n-r-1}(\overline{i})$.

Suggerimento. La quantità al secondo membro di (*b*) è la somma di tutti i prodotti del tipo $p_k(\omega_{j_1} \cdots \omega_{j_{n-r-1}})$, con $j_1 \neq \cdots \neq j_{n-r-1} \neq k$ indici in $I_N \setminus \{i\}$.

d. $D_m(\overline{i}) = D_m - \omega_i D_{m-1}(\overline{i})$.
e. $g(n, r, i) = g(n, r+1, i) + r(1-p_i) D_{n-r}$; $r = 0, 1, \ldots$

Suggerimento. Usare *c* e *d*.

f. Usando la condizione iniziale $g(n, n, i)$, verificare che vale la relazione $g(n, r, i) = (1-p_i) \sum_{t=r}^n t D_{n-r}$.
g. Verificare la relazione (15.20).

Suggerimento. Aggiungere alla popolazione un'unità fittizia $N+1$ con $p_{N+1} = 0$, e osservare che $\sum_{\mathbf{s}\in\mathcal{C}_{N,n}} p(\mathbf{s}) = C_s\, g(n, 0, N+1)$. Applicare quindi la relazione al punto *f*.

15.6. (Probabilità di inclusione del primo ordine nel disegno di Sampford) Provare la (15.21).

Suggerimento. Usando le grandezze introdotte nell'Esercizio 15.5, osservare che $\pi_i = C_s\, \omega_i\, g(n, 1, i)$.

15.7. (Probabilità di inclusione del secondo ordine nel disegno di Sampford) Si consideri la quantità

$$\Psi_{ij} = \sum_{\mathbf{c}\in\mathcal{C}_{N-2,n-2}(\overline{i},\overline{j})} \left\{\prod_{k=1}^{N} \omega_k^{\delta(k;\,\mathbf{c})}\right\} \left(n - p_i - p_j - \sum_{l=1}^{N} p_l \delta(l;\,\mathbf{c})\right)$$

e si supponga di aggregare le due unità i, j in un'unica unità α. La numerosità della risultante popolazione è $N-1$, e si ha $p_\alpha = p_i + p_j$, $\omega_\alpha = p_\alpha/(1-p_\alpha)$. Siano inoltre g', D' definite rispettivamente come g, D nell'Esercizio 15.5, ma relativamente alla nuova popolazione di $N-1$ unità. Provare le seguenti relazioni.

a. $\Psi_{ij} = g'(n,\, 2,\, \alpha) + p_\alpha D'_{n-2}(\overline{\alpha})$.
b. $D'_{n-t} = D'_{n-t}(\overline{\alpha}) + \omega_\alpha D'_{n-t-1}(\overline{\alpha})$ per $t < n-1$ (e con $D'_0(\overline{\alpha}) = 1$).
c. $\Psi_{ij} = \sum_{t=2}^{n}(t - p_\alpha) D'_{n-t}(\overline{\alpha}) = \sum_{t=2}^{n}(t - p_i - p_j) D_{n-t}(\overline{i},\, \overline{j})$.
d. $\pi_{ij} = C_s\, \omega_i\, \omega_j\, \Psi_{ij}$.
e. Sulla base delle $(a) - (d)$, provare la (15.24).

15.8. Provare la relazione (15.25).

Suggerimento. Dalla (d) dell'Esercizio 15.5 si ha

$$N\, D_m - \sum_{i=1}^{N} D_m(\overline{i}) = \sum_{r=1}^{m} (-1)^{r+1} (\omega_1^r + \cdots + \omega_N^r) D_r.$$

Inoltre, $\sum_{i=1}^{N} D_m(\overline{i})$ è una somma di prodotti del tipo $\omega_{j_1} \cdots \omega_{j_m}$, e la combinazione $\{j_1,\, \ldots,\, j_m\}$ compare (una volta) in $N-n$ insiemi $\mathcal{C}_{N-1,m}(\overline{i})$.

15.9. Provare che nel disegno di Sampford con $n = 2$ vale la relazione $\pi_{ij} \leqslant \pi\, \pi_j$ per $i \neq j$.

15.10. Provare che se $\lambda_1 = \lambda_2 = \cdots = \lambda_N$, il disegno di Pareto si riduce a quello semplice senza ripetizione.

Suggerimento. Le variabili aleatorie $Q_1,\, \ldots,\, Q_N$ sono indipendenti e identicamente distribuite, così che Q_i è uguale all'una o all'altra tra $Q_{1:N},\, \ldots,\, Q_{N:N}$ con la stessa probabilità.

15.11. Verificare la relazione (15.32).

Suggerimento. $Pr(Q_i \leqslant y) = Pr(U_i/(1-U_i) \leqslant \vartheta_i\, y) = Pr(U_i \leqslant \vartheta_i\, y/(1+\vartheta_i\, y)) = \vartheta_i\, y/(1+\vartheta_i\, y)$ se $y \geqslant 0$.

15.12. Siano

- $\mathcal{C}_{N-1,n-1}(\overline{i})$: classe di tutte le combinazioni senza ripetizioni di $n-1$ unità di tutta la popolazione I_N privata di i, $I_N \setminus \{i\}$;
- $\mathcal{C}_{N,n-1}(i)$: classe di tutte le combinazioni senza ripetizioni di $n-1$ unità di I_N contenenti i.

Provare la relazione (15.46).

Suggerimento. Tenere conto che

$$\begin{aligned}\pi_i^{(n)} &= \eta_i\, C_{pc}^{(n)} \sum_{\mathbf{s}_{n-1}\in\mathcal{C}_{N-1,n-1}(\overline{i})} \left(\prod_{k=1}^{N} \eta_k^{\delta(k;\,\mathbf{s}_{n-1})}\right) \\ &= \eta_i\, \frac{C_{pc}^{(n)}}{C_{pc}^{(n-1)}} \left\{ \sum_{\mathbf{s}_{n-1}\in\mathcal{C}_{N,n-1}} \left(C_{pc}^{(n-1)} \prod_{k=1}^{N} \eta_k^{\delta(k;\,\mathbf{s}_{n-1})}\right) \right. \\ &\quad \left. - \sum_{\mathbf{s}_{n-1}\in\mathcal{C}_{N,n-1}(i)} \left(C_{pc}^{(n-1)} \prod_{k=1}^{N} \eta_k^{\delta(k;\,\mathbf{s}_{n-1})}\right) \right\}.\end{aligned}$$

15.13. Si considerino un disegno di Poisson condizionato di numerosità n, definito dai numeri $\tau_1, \ldots, \tau_N$, e un disegno di Pareto, sempre di numerosità n, con $\lambda_1 = \tau_1, \ldots, \lambda_N = \tau_N$. Si indichino rispettivamente con $p_{pc}^{(n)}(\mathbf{s})$ e $p_{pa}^{(n)}(\mathbf{s})$ le probabilità dei campioni rispettivamente nel disegno di Poisson condizionato e in quello di Pareto; si indichino inoltre con $\pi_{pc,i}^{(n)}$, $\pi_{pc,ij}^{(n)}$, $\pi_{pa,i}^{(n)}$, $\pi_{pa,ij}^{(n)}$ le probabilità di inclusione del primo e del secondo ordine. Provare che vale la relazione

$$\pi_{pa,i}^{(n)} = \frac{\sum_{k=1}^{N} g_k\, \pi_{pc,ik}^{(n)}}{\sum_{k=1}^{N} g_k\, \pi_{pc,k}^{(n)}}, \quad i = 1, \ldots, N.$$

Suggerimento. Usare la relazione

$$p_{pa}^{(n)}(\mathbf{s}) = \frac{C_{pa}^{(n)}}{C_{pc}^{(n)}}\, p_{pa}^{(n)}(\mathbf{s}) \left(\sum_{k=1}^{N} g_k\, \delta(k;\, \mathbf{s})\right)$$

con $C_{pa}^{(n)}$, $C_{pc}^{(n)}$ costanti opportune, tali che

$$\frac{C_{pa}^{(n)}}{C_{pc}^{(n)}} = 1 \Big/ \sum_{k=1}^{N} g_k\, \pi_{pc,k}^{(n)}.$$

15.14. Si considerino un disegno di Sampford di numerosità n, definito dai numeri $p_1, \ldots, p_N$, e un disegno di Pareto, sempre di numerosità n, con $\lambda_1 = p_1, \ldots, \lambda_N = p_N$. Si indichino rispettivamente con $p_s(\mathbf{s})$ e $p_{pa}(\mathbf{s})$ le probabilità dei campioni rispettivamente nel disegno di Sampford condizionato e in quello di Pareto. Posto $B = \max((1-p_1)/g_1, \ldots, (1-p_N)/g_N)$, verificare che lo schema di rigetto basato sul generare un campione di Pareto e accettarlo se $U \leqslant \sum_{i=1}^{N}(1-p_i)\,\delta(i;\,\mathbf{s}) / \sum_{i=1}^{N} g_i\,\delta(i;\,\mathbf{s})$ produce un campione di Pareto.

Suggerimento. Se $q(\mathbf{s}) = \prod_{i=1}^{N} p_i^{\delta(i;\,\mathbf{s})}\,(1-p_i)^{1-\delta(i;\,\mathbf{s})}$, si ha $p_s(\mathbf{s}) = A_s\, q(\mathbf{s}) \sum(1-p_i)\,\delta(i;\,\mathbf{s})$, e $p_{pa}(\mathbf{s}) = q(\mathbf{s}) \sum g_i\,\delta(i;\,\mathbf{s})$, da cui $q(\mathbf{s}) \sum(1-p_i)\,\delta(i;\,\mathbf{s}) \leqslant B\, q(\mathbf{s}) \sum g_i\,\delta(i;\,\mathbf{s})$ se $\sum(1-p_i)\,\delta(i;\,\mathbf{s}) \leqslant B \sum g_i\,\delta(i;\,\mathbf{s})$, il che accade se $B \geqslant (1-p_i)/g_i$ per ciascun $i = 1, \ldots, N$.

15.15. Provare la relazione (15.51).

15.16. Provare la relazione (15.53).

15.17. Provare la relazione (15.58).

15.18. Dato un disegno campionario di Bernoulli con dimensione campionaria attesa pari a $Np = n$, si consideri lo stimatore di Horvitz-Thompson della media della popolazione. Mostrare che l'effetto del disegno è pari a

$$deff(B,\ t_{HT}) \approx \left(1 + \frac{1}{CV_y^2}\right) \tag{15.75}$$

dove CV_y è il coefficiente di variazione di Y nella popolazione.

Suggerimento. Utilizzare la relazione

$$\sum_{i=1}^{N} y_i^2 = NS_y^2 \left(1 - \frac{1}{N} + \frac{1}{CV_y^2}\right). \tag{15.76}$$

Bibliografia

Basu, D.: An essay on the logical foundations of survey sampling. Part 1. In: Godambe, V. P., Sprott, D. A. (eds.) Foundations of Statistical Inference. Holt, Rinehart & Winston, Toronto (1971)

Bondesson, L.: Recursion Formulas for Inclusion Probabilities of All Orders for Conditional Poisson, Sampford, Pareto, and More General Sampling Designs. In: Carlson, M., Nyquist, H., Villani, M. (eds.) Official Statistics – Methodology and Applications in Honour of Daniel Thorburn. Brommatryck and Brolins AB, Stoccolma (2010)

Bondesson, L., Traat, I., Lundqvist, A.: Pareto sampling versus Sampford and conditional Poisson sampling. Scandinavian Journal of Statistics **33**, 699–720 (2006)

Brewer, K. R. W.: A simple procedure for πpswor. Australian Journal of Statistics **17**, 166–172 (1975)

Brewer, K. R. W., Hanif M.: Sampling with Unequal Probabilities. Springer Verlag, New York (1983)

Cassel, C. M., Särndal, C.-E., Wretman, J.: Foundations of Inference in Survey Sampling. Wiley, New York (1977)

Chaudhuri, A., Vos, J. W. E.: Unified Theory and Strategies of Survey Sampling. North-Holland, Amsterdam (1988)

Chauvet, G., Tillé, Y.: A fast algorithm for balanced sampling. Computational Statistics **21**, 53–62 (2006)

Cicchitelli, G., Herzel, A., Montanari, G. E.: Il campionamento statistico. Il Mulino, Bologna (1992)

Cochran, W. G.: Comparison of methods for determining stratum boundaries. Bulletin of the International Statistical Institute **38**, 345–358 (1961)

Cochran, W. G.: Sampling Techniques, 3a ed. Wiley, New York (1977)

Connor, W. S.: An exact formula for the probability that specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. Journal of the American Statistical Association **61**, 384–490 (1966)

Cox, D. R., Hinkley, D. V.: Theoretical Statistics. Chapman & Hall, Londra (1974)
Dalenius, T.: The problem of optimum stratification. Skandinavisk Aktuarietidskrift **33**, 203–213 (1950)
Dalenius, T., Hodges, J. L.: Minimum variance stratification. Journal of the American Statistical Association **54**, 88–101 (1959)
Deville, J.-C., Tillé, Y.: Unequal probability sampling without replacement through a splitting method. Biometrika **85**, 89–101 (1998)
Deville, J.-C., Tillé, Y.: Efficient balance sampling: the cube method. Biometrika **91**, 893–912 (2004)
Deville, J.-C., Tillé, Y.: Variance approximation under balanced sampling. Journal of Satistical Planning and Inference **128**, 411–425 (2005)
Ekman, G.: An approximation useful in univariate stratification. The Annals of Mathematical Statistics **30**, 219–229 (1959)
Gini, C., Galvani, L.: Di un'applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione (1 dicembre 1921). Annali di Statistica, Serie VI, **4**, 1–107 (1929)
Godambe, V. P., Joshi, V. M.: Admissibility and bayes estimation in sampling finite populations i. Annals of Mathematical Statistics **36**, 1707–1722 (1965)
Golder, P. A., Yeomans, K. A.: The use of cluster analysis for stratification. Applied Statistics **22**, 213–219 (1973)
Hájek, J.: Limiting distributions in simple eandom sampling from a finite population. Publications of the Mathematical Institute Hungarian Academy of Sciences **5**, 361–374 (1960)
Hájek, J.: Comments on "An essay on the logical foundations of survey sampling. Part 1." In: Godambe, V. P., Sprott, D. A. (eds.) Foundations of Statistical Inference. Holt, Rinehart & Winston, Toronto (1971)
Hájek, J.: Sampling from a finite popolation. Marcel Dekker, New York (1981)
Hedlin, D.: A procedure for stratification by an extended Ekman rule. Journal of Official Statistics **16**, 15–29 (2000)
Herzel, A.: On mean values and unbiased estimators in simple random sampling. Statistica, 315–350 (1982)
Hess, I., Sethi, V. K., Balakrishnan, T. R.: Stratification: A practical investigation. Journal of the American Statistical Association (1966)
Hidiroglou, M. A., Srinath, K. P.: Problems associated with designing subannual surveys. Journal of Business & Economic Statistics **11**, 397–404 (1993)
Horvitz, D. G., Thompson, D. J.: A generalisation of sampling without replacement from a finite universe. Journal of the American Statistical Association **47**, 663–685 (1952)
Kish, L.: Survey Sampling. Wiley, New York (1965)
Lavallée, P., Hidiroglou, M. A.: On the stratification of skewed populations. Survey Methodology **14**, 33–43 (1988)

Madow, W. G.: On the theory of systematic sampling, ii. Annals of Mathematical Statistics **20**, 333–354 (1949)

Matei, A., Tillé, Y.: Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. Journal of Official Statistics **21**, 543–570 (2005)

Murthy, M. N.: Sampling Theory and Methods. Statistical Publishing Society, Calcutta (1967)

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

Rao, J. N. K.: Ratio and regression estimators. In: Johnson, N. L., Smith, H. (eds.) New Developments in Survey Sampling. Wiley, New York (1969)

Rivest, L.-P.: A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. Survey Methodology **28**, 191–198 (2002)

Rosén, B.: Asymptotic theory for order sampling. Journal of Statistical Planning and Inference **62**, 135–158 (1997a)

Rosén, B.: On sampling with probability proportional to size. Journal of Statistical Planning and Inference **62**, 159–191 (1997b)

Royall, R. M., Herson, J.: Robust estimation in finite populations i. Journal of the American Statistical Association **68**, 880–889 (1973)

Sampford, M. R.: On sampling without replacement with unequal probabilities of selection. Biometrika **54**, 499–513 (1967)

Särndal, C. E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer Verlag, New York (1993)

Serfling, R. J.: Approximately optimum stratification. Journal of the Americal Statistical Association **63**, 1298–1309 (1968)

Singh, R.: Approximately optimum stratification on the auxiliary variable. Journal of the Americal Statistical Association **66**, 829–833 (1971)

Slanta, J., Krenzke, T.: Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the census bureau's annual capital expenditures surveys. Survey Methodology **22**, 65–7 (1990)

Stein, C.: Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 197–206 (1956)

Tillé, Y.: Sampling Algorithms. Springer Verlag, New York (2006)

Tillé, Y., Matei, A.: Sampling: Survey sampling. R package version 2.2. http://cran.r-project.org/src/contrib/descriptions/sampling.html (2009)

Wolter, K. M.: Introduction to Variance Estimation, 2a ed. Springer Verlag, New York (2007)

Yates, F., Grundy, P. M.: Selection without replacement from within strata with probability proportional to sise. Journal of the Royal Statistical Society B **35**, 235–261 (1953)

Indice analitico

Unitext – Collana di Statistica e Probabilità Applicata

C. Rossi, G. Serio
La metodologia statistica nelle applicazioni biomediche
1990, XVIII, 354 pp, ISBN 978-3-540-52797-8

A. Azzalini
Inferenza statistica
Una presentazione basata sul concetto di verosimiglianza
2a ed., 2001, XIV, 367 pp, ISBN 978-88-470-0130-5

E. Bee Dagum
Analisi delle serie storiche: modellistica, previsione e scomposizione
2002, XII, 302 pp, ISBN 978-88-470-0146-6

B. Luderer, V. Nollau, K. Vetters
Formule matematiche per le scienze economiche
2003, X, 212 pp, ISBN 978-88-470-0224-1

A. Azzalini, B. Scarpa
Analisi dei dati e data mining
Corr. 2a ed., 2004, X, 232 pp, ISBN 978-88-470-0272-2

F. Battaglia
Metodi di previsione statistica
2007, X, 323 pp, ISBN 978-88-470-0602-7

L. Piccinato
Metodi per le decisioni statistiche
2a ed., 2009, XIV, 474 pp, ISBN 978-88-470-1077-2

E. Stanghellini
Introduzione ai metodi statistici per il credit scoring
2009, X, 177 pp, ISBN 978-88-470-1080-2

A. Rotondi, P. Pedroni, A. Pievatolo
Probabilità, Statistica e Simulazione
3a ed., 2011, XIV, 540 pp, ISBN 978-88-470-2363-5

P. L. Conti, D. Marella
Campionamento da popolazioni finite. Il disegno campionario
2012, XIV, 444 pp, ISBN 978-88-470-2576-9

La versione online dei libri pubblicati nella serie è disponibile
su SpringerLink. Per ulteriori informazioni, visitare il sito:
http://www.springer.com/series/1380