

3- 2005



SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

TECNICHE E STRUMENTI

GENESEES V. 3.0

Funzione Stime ed Errori

*Manuale utente
e aspetti metodologici*



 Istat



SISTEMA STATISTICO NAZIONALE
ISTITUTO NAZIONALE DI STATISTICA

GENESEES V. 3.0

Funzione Stime ed Errori

*Manuale utente
e aspetti metodologici*

A cura di: Daniela Pagliuca
e-mail: pagliuca@istat.it

Sezione I: Cap. 1, Cap. 2, Cap. 3 Daniela Pagliuca; Cap. 4 Stefano Falorsi; Cap. 5 Daniela Pagliuca e Patrizia Giaquinto; Cap. 6 Paolo Righi; Cap. 7 Daniela Pagliuca.

Sezione II: Cap. 1: paragrafi 1.1 e 1.2 Daniela Pagliuca, paragrafi 1.3 e 1.4 Paolo Righi (in particolare i sottoparagrafi 1.3.1.3, 1.3.1.4, e i sottoparagrafi 1.3.2.1, 1.3.2.2, 1.3.2.3, 1.3.2.4, 1.3.2.5, 1.3.2.6, 1.3.2.7 sono ad opera di Paolo Righi e Fabrizio Solari); Cap. 2: paragrafi 2.1 e 2.2 Daniela Pagliuca, paragrafi 2.3 e 2.4 Stefano Falorsi, paragrafo 2.5 Daniela Pagliuca;

Sezione III: Daniela Pagliuca, Loredana Di Consiglio;

Appendici: A.1 Paolo Righi; A.2 Fabrizio Solari; A.3 Stefano Falorsi; A.4 Paolo Righi; A.5 Stefano Falorsi.

GENESEES V. 3.0

Funzione Stime ed Errori

Manuale utente e aspetti metodologici

Istituto nazionale di statistica
Via Cesare Balbo, 16 - Roma

Coordinamento editoriale:
Piero Crivelli
Servizio Produzione editoriale
Via Tuscolana, 1788 - Roma

Progetto grafico e videoimpaginazione:
Antonio Maggiorani

Stampa digitale:
Istat - Produzione libraria e centro stampa

Luglio 2005 – copie 250

Si autorizza la riproduzione ai fini
non commerciali e con citazione della fonte

GENESEES V 3.0

(GENERALised software for Sampling Estimates and Errors in Surveys)

Software generalizzato per il calcolo dei pesi, delle stime e degli errori campionari

Genesees V. 3.0 è un software generalizzato nato da diverse procedure SAS, sviluppate da Piero Demetrio Falorsi e Stefano Falorsi, per il calcolo dei pesi e delle stime mediante stimatori di regressione generalizzata, per il calcolo degli errori campionari, per la loro presentazione sintetica mediante modelli regressivi. Tali procedure, dal punto di vista dell'architettura e degli algoritmi utilizzati, costituiscono la base delle funzioni di "Riponderazione" e di "Stima ed Errori campionari" attualmente disponibili anche in Genesees V. 3.0; rispetto alla versione 2.0 il software Genesees V. 3.0 comprende una funzione aggiuntiva, la funzione Analisi dei Modelli, che agevola l'utente nella rappresentazione sintetica degli errori campionari, permettendo la visualizzazione grafica dei dati per tenere in considerazione e eventualmente eliminare i valori estremi. Genesees V. 3.0 è stato realizzato all'interno di un progetto di sviluppo dell'unità MTS/F "Software generalizzati per la produzione statistica" dell'Istat, responsabile Daniela Pagliuca, in collaborazione con l'unità PSM / A "Strategia campionaria e tecnica di rilevazione", responsabile Stefano Falorsi. Il progetto ha avuto come obiettivo quello di ottimizzare le procedure SAS, implementando i controlli necessari per l'esecuzione e sviluppando una interfaccia user-friendly per consentire agli utenti un'interazione di tipo avanzato, e di implementare ex-novo la funzione Analisi Modelli. Stefano Falorsi è il responsabile delle metodologie statistiche implementate nel software. Si ringraziano Piero Falorsi e Giulio Barcaroli per i commenti ed i suggerimenti.

Indice

Presentazione	9
 SEZIONE I:	
IL SOFTWARE GENESEES V. 3.0 E LA FUNZIONE DI CALCOLO DELLE STIME E DEGLI ERRORI CAMPIONARI	
1. Introduzione: il contenuto del manuale	15
1.1 Cosa contiene il manuale	15
1.2 Come utilizzare il manuale: alcune indicazioni sui capitoli	16
2. L'installazione e l'avvio del software	23
2.1 I requisiti hardware e software e modalità di installazione	23
2.2 La procedura di avvio e la password di esecuzione	25
2.3 Assistenza al software	26
3. Il software Genesees V. 3.0: un insieme di funzioni	31
3.1 La struttura del software Genesees V. 3.0	31
3.2 Le funzioni del software Genesees V. 3.0	34
4. La funzione Stime ed Errori campionari: cenni metodologici	37
5. L'utilizzo della funzione di calcolo delle Stime degli Errori di campionamento del software Genesees V. 3.0	45
5.1 La schermata principale	45
5.2 Il calcolo delle stime e degli errori campionari	47
5.2.1 <i>Le variabili e i parametri di input</i>	50
5.2.2 <i>La selezione delle variabili di input tramite la maschera di selezione</i>	52

5.2.3 <i>La selezione delle variabili di input tramite i parametri attivati dal software</i>	57
5.2.4 <i>L'elaborazione</i>	58
5.3 La funzione "Crea stampe"	60
6. La descrizione delle stampe	71
6.1 Stampa 1	71
6.2 Stampa 2	72
6.3 Stampa 3	74
6.4 Stampa 4	75
6.5 Stampa 5	77
6.6 Stampa 6	78
6.6 Stampa 7	79
6.6 Stampa 8	81
7. I file di output della funzione di Stime ed Errori di Genesees	85

SEZIONE II:

Approfondimenti sulla costruzione dell'input e sui data-set di output della funzione di calcolo delle Stime e degli Errori di Genesees V. 3.0

1. La costruzione del data-set di input	91
1.1 Le variabili ed i parametri di input	91
1.1.1 <i>Le variabili di input</i>	92
1.1.2 <i>I parametri di input</i>	97
1.2 I vincoli sulle variabili	98
1.3 Definizione delle variabili di input in relazione alla strategia campionaria	102
1.3.1 <i>Definizione delle variabili di input per un dato stimatore</i>	103
1.3.2 <i>Definizione delle variabili di input per un dato disegno</i>	127
1.4 Definizione delle variabili di input per il livello della stima (dominio di stima) considerato	136
1.4.1 <i>Definizione delle variabili di input per i domini di stima pianificati</i>	137
1.4.2 <i>Definizione delle variabili di input per i domini di stima non pianificati</i>	138
2. I data-set di output	141
2.1 Il data-set dei parametri di input	141

2.2 Gli errori rilevati sul data-set di input	142
2.3 I data-set con le informazioni su stime ed errori campionari	144
2.4 I data-set con le informazioni sulla stratificazione e sul campione	155
2.5 I data-set con informazioni per elaborazioni successive e file di output	158

SEZIONE III:

Un esempio di utilizzo della funzione di Stima ed Errori campionari

1. L'applicazione della funzione di Stime ed Errori campionari di Genesees V. 3.0	161
1.1 La costruzione del data-set di input	162
1.1.1 <i>Il data-set di esempio</i>	163
1.1.2 <i>La costruzione delle variabili di input</i>	164
1.2 L'uso del software e la presentazione dell'output	169
1.2.1 <i>L'uso delle schermate utilizzando il data-set di esempio</i>	169
1.2.2 <i>Le stampe che si ottengono utilizzando il data-set di esempio</i>	176

Appendici

A.1 Cenni sulla definizione dello stimatore di regressione generalizzata	189
A.1.1 Gruppo di riferimento del modello	194
A.1.2 Livello del modello	195
A.1.3 Tipo di modello	197
A.2 Linearizzazione dello stimatore di regressione generalizzata	199
A.3 Lo stimatore di regressione generalizzata per i diversi disegni di campionamento	203
A.3.1 Campionamento di unità elementari con probabilità d'inclusione costanti	203
A.3.2 Campionamento a grappoli con probabilità d'inclusione costanti	205
A.3.3 Campionamento di unità elementari con probabilità d'inclusione variabili	207

A.3.4 Campionamento a grappoli con probabilità d'inclusione variabili	208
A.3.5 Campionamento a due o più stadi	209
A.4 La costruzione dei data-set di input per definire i gruppi di riferimento	213
A.4.1 Costruzione dei gruppi di riferimento: caso I	213
A.4.2 Costruzione dei gruppi di riferimento: caso II	226
A.5 Presentazione sintetica degli errori di campionamento mediante modelli regressivi	231
A.5.1 Introduzione	231
A.5.2 Caratteristiche generali del metodo	233
A.5.3 Il caso delle stime di frequenze	238
A.5.4 Il caso delle stime di totali di variabili quantitative	243
Bibliografia	247

Presentazione

Gran parte delle rilevazioni compiute dagli Istituti nazionali di statistica sono effettuate mediante l'osservazione, anziché dell'intera popolazione di interesse, di suoi sottoinsiemi, scelti con criteri rigorosamente scientifici, tali da massimizzare il rapporto tra l'accuratezza delle stime prodotte ed i costi di rilevazione.

Le stime calcolate mediante i dati campionari costituiscono l'obiettivo fondamentale delle rilevazioni: il soddisfacimento della domanda informativa relativa ai fenomeni oggetto di studio. La produzione di tali stime, e la loro diffusione, non esauriscono però il compito dello statistico responsabile dell'indagine.

È ormai pratica corrente di tali Istituti, e, più in generale, degli enti produttori di statistiche ufficiali, quella di fornire, assieme ai valori puntuali delle stime prodotte mediante le varie indagini, anche indicazioni riguardanti la loro accuratezza, intesa come "vicinanza" tra i valori veri e quelli stimati.

Schematicamente, l'accuratezza delle stime dipende, da una parte, dalla presenza degli errori non campionari (errori di copertura, di mancata risposta totale o parziale, di misura e di elaborazione), dall'altra, dagli errori campionari. I primi sono dovuti, sostanzialmente, ad imperfezioni del sistema di raccolta e trattamento dei dati, mentre i secondi si riferiscono all'incertezza dovuta al fatto che solo un sottoinsieme di unità della popolazione è sottoposto a rilevazione, anziché l'intera popolazione oggetto di studio.

Mentre la valutazione della prima tipologia di errori comporta generalmente il ricorso a fonti esterne di dati, oppure a rilevazioni aggiuntive

(quali, ad esempio, le indagini di copertura e di qualità), che possono essere anche molto onerose, al contrario la valutazione degli errori campionari può essere condotta direttamente sui dati osservati, senza ulteriori costi da sopportare per i responsabili delle indagini.

Gli errori campionari prodotti da campioni casuali semplici possono essere facilmente calcolati. Al contrario, qualora si considerino strategie campionarie che prevedano disegni complessi (presenza di più stadi, stratificazione delle unità, schemi di selezione a probabilità variabile, ecc.) e stimatori non lineari (quali ad esempio quelli che fanno uso di tecniche di calibrazione), allora il calcolo della variabilità delle stime dovuta agli errori campionari non può certo dirsi banale.

Strategie campionarie di crescente complessità contraddistinguono le rilevazioni effettuate dai produttori di statistiche ufficiali, e tra questi l'Istat in modo particolare. È questo il motivo, unitamente a considerazioni legate al software esistente sul mercato, che ha spinto l'Istituto a dotarsi di un sistema, sviluppato in proprio, che permette di raggiungere l'obiettivo per la generalità delle indagini condotte.

Il software per il calcolo degli errori campionari nasce dallo studio e dalle attività di alcuni ricercatori del Servizio Studi Metodologici dell'Istat negli anni '80. Il Servizio Studi, anche in quegli anni, garantiva la copertura delle fasi peculiari delle indagini campionarie: da un lato, la progettazione del campione (definizione della dimensione, degli strati, allocazione delle unità, scelta delle modalità di selezione), dall'altro l'elaborazione delle stime (con un eventuale preventivo passo di riponderazione dei dati nel caso di utilizzo di stimatori di calibrazione) e la valutazione del grado di affidabilità di queste. In particolare, uno degli obiettivi era di disporre di strumenti che permettessero di coprire integralmente questa seconda fase. Per primo, fu sviluppato un prototipo software per calcolare i pesi campionari finali tenendo conto di totali noti della popolazione oggetto di studio e garantendo la coincidenza tra questi e le corrispondenti stime campionarie. Immediatamente dopo venne implementato un secondo prototipo per calcolare le stime e gli errori campionari. I due prototipi sono stati sviluppati da Piero Falorsi e Stefano Falorsi (Falorsi P. e Falorsi S., 1995; Falorsi P. e Falorsi S., 1997), attualmente dirigenti del Servizio *Progettazione e Supporto Metodologico nei processi di produzione statistica (PSM)* dell'ISTAT. La disponibilità di tali strumenti permise di trattare in modo efficace ed omogeneo le

fasi di elaborazione dei dati campionari, relativamente alle più importanti indagini condotte dall'Istituto. Con un limite, però: le caratteristiche dei due sistemi, dal punto di vista di facilità di utilizzo, non erano tali da permettere un uso agevole ad utenti che non fossero quelli già esperti del Servizio Studi. Data la scarsità di risorse in tale Servizio, il trattamento di più indagini in parallelo è stato spesso difficoltoso. Nell'ottica poi di estendere l'applicazione delle tecniche di calibrazione e di valutazione della varianza campionaria anche alla più vasta utenza potenziale del SISTAN, si comprende come questo limite diventasse difficilmente accettabile.

Per tale motivo, ed anche al fine di ottimizzare l'efficienza elaborativa degli algoritmi interni, si decise di sviluppare software di uso generale, partendo dai suddetti prototipi. Si scelse di procedere con lo sviluppo interno – anziché utilizzare procedure statistiche disponibili presso altri enti statistici o prodotti di mercato – per due motivi: da un lato, assicurare al software le stesse caratteristiche metodologiche già implementate nei prototipi, di cui era nota la capacità di soddisfare le esigenze della quasi totalità delle indagini ISTAT, caratterizzate da un'alta complessità delle strategie campionarie adottate. Dall'altro, garantirsi la possibilità di poter intervenire in qualsiasi momento ed in piena autonomia al fine di arricchire i sistemi con le tecniche innovative che la ricerca costantemente produce in questo settore.

Si costituì quindi un gruppo, composto, oltre che dagli autori del prototipo di calcolo degli errori campionari, anche da Daniela Pagliuca e Germana Scepi, ai fini di ottimizzarne le prestazioni e garantirne la generalizzazione. La prima versione del nuovo software è stata presentata a Praga al convegno ETK'99 – Exchange of Technology and Knowledge – 1999 (Falorsi, Pagliuca e Scepi, 1999; Falorsi, Pagliuca e Scepi, 2000).

In seguito alla nascita dell'unità che si occupa di software generalizzato per la produzione statistica (attualmente, collocata nel Servizio *Metodologie, Tecnologie e Software* con la denominazione MTS/F - “Software generalizzati per la produzione statistica”), la cui responsabilità è stata affidata a Daniela Pagliuca e che ha previsto l'inserimento nel progetto di informatici esperti quali Roberto Di Giuseppe e Marco Landriscina è stato possibile realizzare le successive versioni, caratterizzate da una sempre maggiore integrazione di funzionalità.

La prima versione, Genesees V. 1.0, (presentato a Berlino al convegno

Compstat 2002, Pagliuca e Righi, 2002) conteneva la sola funzione di calcolo delle stime e della varianza campionaria. La versione successiva (Genesees v2.0) unificava in un unico sistema le due funzionalità per il calcolo delle stime e della varianza campionaria, da una parte, e per la riponderazione delle osservazioni campionarie, dall'altra.

L'attuale versione del software – Genesees V. 3.0 – garantisce, oltre alle due funzioni citate sopra, anche quella per la stima e l'analisi dei modelli per la presentazione degli errori campionari, funzione implementata ex-novo nell'ambito di un progetto diretto dall'unità MTS/F, per agevolare l'utente nella rappresentazione sintetica degli errori campionari, permettendo la visualizzazione grafica dei dati per individuare, ed eventualmente eliminare, i valori *estremi*.

La funzione cui questo manuale si riferisce è quella di **Calcolo delle Stime e degli errori**, contenuta in Genesees V. 3.0 (GENeralised Sampling Estimates and Errors in Surveys).

Giulio Barcaroli

*Responsabile del Servizio
Metodologie, tecnologie e
software per la produzione statistica*

Piero Demetrio Falorsi

*Responsabile Servizio
Progettazione
e Supporto Metodologico*

SEZIONE I

**I software Genesee V. 3.0
e la funzione di calcolo delle stime
e degli errori campionari**

1. Introduzione: il contenuto del manuale

Il presente manuale guida gli utenti che devono fare uso della **FUNZIONE DI STIME ED ERRORI CAMPIONARI** del software **Genesees V. 3.0**.

In particolare:

- Aiuta l'utente ad installare il software Genesees V. 3.0, evidenziando i requisiti hardware e software richiesti;
- Descrive la struttura del software Genesees V. 3.0 nel suo complesso, come insieme di funzioni;
- Descrive la metodologia che è alla base della funzione di Stime ed Errori del software Genesees V. 3.0;
- Presenta la funzione di Stime ed Errori, descrivendo le maschere che possono essere richiamate per il calcolo delle stime e degli errori campionari;
- Descrive come costruire l'input appropriato per il calcolo delle stime e degli errori campionari e analizza i dati di output;
- Illustra le stampe ottenibili tramite la funzione di Stime ed Errori;
- Presenta un esempio di applicazione.

1.1 Cosa contiene il manuale

Il manuale è diviso in tre sezioni e comprende inoltre delle appendici metodologiche.

La **Sezione I** costituisce il manuale vero e proprio per l'**utilizzo della funzione di Stime ed Errori**: in essa è descritta la base metodologica, vengono illustrate le schermate presentate dal software e le stampe.

Gli approfondimenti relativi ai dati di input e di output vengono demandati alla sezione successiva.

I primi tre capitoli sono introduttivi al software Genesees V. 3.0: il presente *capitolo 1* illustra il contenuto del manuale e la modalità di utilizzo; il *capitolo 2* aiuta l'utente ad installare ed avviare il software e il *capitolo 3* si riferisce al software nel suo complesso.

Dopo i primi tre capitoli introduttivi al software, il manuale descrive in dettaglio la funzione di Stime ed Errori.

La **Sezione II** approfondisce i dati di input e output della funzione di Stime ed Errori, illustrando dettagliatamente come costruire il data-set di input e descrivendo i data-set di output. È da osservare che per utilizzare la funzione di Stime ed Errori è richiesta la costruzione di un data-set di input e tale operazione deve effettuarsi seguendo criteri ben definiti. La configurazione del data-set di input è perciò trattata come approfondimento nella Sezione II, in quanto è rivolta a chi, avendo una adeguata preparazione metodologica, è in grado di comprendere le scelte sottostanti il campione. Anche i data-set di output sono approfonditi in questa seconda sezione.

La **Sezione III** descrive un esempio in cui si illustra come calcolare le stime e gli errori campionari nel caso di una applicazione che è stata costruita ad hoc per mostrare quanto descritto nelle due parti precedenti.

Le **APPENDICI** approfondiscono gli aspetti metodologici alla base della funzione di calcolo delle stime e degli errori campionari del software Genesees V. 3.0.

1.2 Come utilizzare il manuale: alcune indicazioni sui capitoli

Per agevolare l'utilizzo del software vengono di seguito riportate alcune indicazioni utili per l'utente, descrivendo quanto riportato nei capitoli del manuale.

La Sezione I è formata dai Capitoli 1, 2, 3, 4, 5, 6, 7.

La Sezione 2 è formata dai Capitoli 1, 2.

La Sezione 3 non è suddivisa in capitoli.

All'interno del manuale, il richiamo ad altri capitoli o paragrafi (quale ad esempio: *cfr. paragrafo 4.3*), ove non venga specificata la sezione di riferimento, va inteso riferito alla stessa sezione.

SEZIONE I

Nella **Sezione I** è possibile leggere come utilizzare la funzione del software Genesees V. 3.0 per il calcolo delle stime e degli errori campionari.

Nel dettaglio un utente può leggere come:

- a) Installare il software
- b) Utilizzare le schermate presentate dal software
- c) Selezionare le stampe desiderate e leggere i dati di output

Capitolo 1

Il presente *capitolo 1* è introduttivo e illustra il contenuto del manuale e il suo utilizzo.

Capitolo 2

Il *capitolo 2* descrive la procedura di installazione ed avvio del software Genesees V. 3.0.

Per **installare il software** l'utente riceve un CD-ROM contenente un programma di installazione, le cui informazioni essenziali sono riportate nel *capitolo 2*.

Tali informazioni sono anche disponibili (*aggiornamento 2005*):

- via internet (per utenti esterni all'Istat):
<http://www.istat.it/Metodologi/index.htm> (selezionare "Metodi e Software per indagini statistiche").
- via intranet (per utenti Istat).

Capitolo 3

Dopo l'installazione è utile leggere il *capitolo 3*, introduttivo a Genesees V. 3.0. Il *capitolo 3* infatti presenta il software Genesees V. 3.0 nel suo complesso, come **insieme di funzioni**.

Capitolo 4

Il *capitolo 4* introduce i **cenni metodologici** alla base della funzione di Stime ed Errori del software Genesees V. 3.0.

Capitolo 5

Il *capitolo 5* descrive in dettaglio **come usare le schermate del software Genesees V. 3.0 per il calcolo delle stime e degli errori campionari** e presenta sommariamente le stampe, approfondite nel successivo *capitolo 6*.

I *paragrafi 5.1 e 5.2* supportano l'utente descrivendo come utilizzare le maschere del software; il *paragrafo 5.3* illustra come produrre le stampe.

In dettaglio:

Il *paragrafo 5.1* è introduttivo e indica come avviare il software, riprendendo quanto già descritto nel capitolo 3 (in riferimento a Genesees V. 3.0, visto nella sua globalità come insieme di funzioni).

Il *paragrafo 5.2* entra nel merito della descrizione dell'uso della funzione di Stime e di Errori Campionari: il *paragrafo 5.2.1* introduce le variabili e i parametri di input per la funzione di Stime ed Errori; nei *paragrafi 5.2.2 e 5.2.3* è descritto come selezionare tali variabili di input; il *paragrafo 5.2.4* illustra come eseguire l'elaborazione vera e propria per ottenere le stime ed errori campionari.

L'utente può selezionare le informazioni che desidera ottenere in stampa, scegliendo tra otto possibili output: nel *paragrafo 5.3* viene descritta la selezione delle diverse tabelle e sono indicate le informazioni che è possibile ottenere.

Capitolo 6

Le **informazioni contenute nelle stampe** create dalla funzione di Stime ed Errori sono approfondite nel *capitolo 6*.

Capitolo 7

Nel *capitolo 7* viene descritto il tipo di output ottenibile dalla funzione di Stime ed Errori, in termini di file e *data-set*, soffermandosi in particolare sui **file** di output: infatti il software permette di memorizzare le stampe su file *ascii* ed *excel*.

SEZIONE II

Nella **Sezione II** è possibile leggere gli approfondimenti sull'input e output della funzione del software Genesees V. 3.0 per il calcolo delle stime e degli errori campionari.

Nel dettaglio l'utente può leggere come:

- 1) Costruire l'input
- 2) Capire le informazioni contenute nei *data-set* di output e eventualmente utilizzare questi ultimi per altre elaborazioni.

Capitolo 1

Nel *capitolo 1* è possibile approfondire **la costruzione del data-set di input** e, in particolare, si illustra come costruire le variabili sulla base del campione (tipo di stimatore, disegno etc.).

E' necessario:

1) **Predisporre l'input**

Per costruire l'input, l'utente deve essere a conoscenza delle variabili di input da creare e dei parametri richiesti dal software: le variabili e i parametri di input sono descritti nel *paragrafo 1.1*.

Nel *paragrafo 1.2* sono presentati i vincoli che tali variabili devono rispettare.

2) **Definire le variabili di input sulla base del tipo di stimatore utilizzato** (*paragrafo 1.3.1*)

L'utente deve definire le variabili del *data-set* di input in base al gruppo di riferimento del modello (per far ciò è utile leggere i *paragrafi* 1.3.1.1, 1.3.1.2); definire le variabili del *data-set* di input in base al livello di unità considerate, unità elementari o grappolo (per far ciò è utile leggere il *paragrafo* 1.3.1.3); definire le variabili del *data-set* di input in base al tipo di modello (per far ciò è utile leggere il *paragrafo* 1.3.1.4).

3) **Definire le variabili di input sulla base del tipo di disegno campionario sottostante l'indagine** (*paragrafo* 1.3.2).

L'utente deve definire le variabili del *data-set* di input in base al disegno campionario che è stato adottato quando sono stati calcolati i coefficienti finali di input. In particolare le modalità di costruzione del *data-set* di input sono presentate prendendo in considerazione i seguenti possibili disegni campionari:

nel *paragrafo* 1.3.2.1 è esaminato il campionamento stratificato di unità elementari con reimmissione e con probabilità di selezione nel campione costante;

nel *paragrafo* 1.3.2.2 è esaminato il campionamento stratificato di grappoli di unità elementari con reimmissione e con probabilità di selezione nel campione costante;

nel *paragrafo* 1.3.2.3 è esaminato il campionamento stratificato di unità elementari senza reimmissione e con probabilità di inclusione nel campione costante;

nel *paragrafo* 1.3.2.4 è esaminato il campionamento stratificato di grappoli di unità elementari senza reimmissione e con probabilità di inclusione nel campione costante;

nel *paragrafo* 1.3.2.5 è esaminato il campionamento stratificato di unità elementari con o senza reimmissione e con probabilità di inclusione nel campione variabile;

nel *paragrafo* 1.3.2.6 è esaminato il campionamento stratificato di grappoli di unità elementari con o senza reimmissione e con probabilità di inclusione nel campione variabile;

nel *paragrafo* 1.3.2.7 è esaminato il campionamento a due o più stadi di selezione.

4) **Definire le variabili di input sulla base del dominio di stima**

Alcune variabili di input definiscono il livello della stima che si desidera ottenere; le stime possono essere calcolate per domini pianificati (*paragrafo 1.4.1*) e per domini non pianificati (*paragrafo 1.4.2*).

Capitolo 2

Nel *capitolo 2* della *Sezione II* sono illustrati dettagliatamente i *data-set* di output del software, relativi all'utilizzo della *funzione di calcolo delle stime e degli errori campionari*.

Alcuni *data-set* sono utili anche per successive elaborazioni.

Tra i vari *data-set* di output, un *data-set* è creato dall'elaborazione per memorizzare parametri di input (*paragrafo 2.1*), un secondo *data-set* memorizza eventuali errori rilevati sull'input (*paragrafo 2.2*); altri *data-set* contengono le informazioni sulla stratificazione, sul campione (*paragrafo 2.4*) e sulle stime ed errori campionari (*paragrafo 2.3* e *2.5*).

SEZIONE III

La Sezione IIII infine descrive una applicazione della funzione di calcolo delle stime e degli errori campionari, ripercorrendo i passi descritti nei capitoli precedenti.

L'utente può utilizzare il *data-set* SAS di esempio *esempio.sas7bdat* per sperimentare il software e comprendere al meglio quanto indicato in questa sede. Tale *data-set* è costruito ad hoc per l'applicazione ed è utilizzabile dopo l'installazione (è memorizzato nella cartella *c:\genesees\Esempi*). Vengono inoltre commentati alcuni risultati.

Ogni capitolo del manuale - e paragrafo ove necessario - è introdotto da una sintesi che aiuta l'utente ad orientarsi nell'uso del manuale stesso.

Per chiarimenti sull'utilizzo del manuale e del software si può utilizzare l'indirizzo di posta elettronica mts-f@istat.it.

2. L'installazione e l'avvio del software

***Sintesi:** In questo capitolo vengono riportati i requisiti hardware e software richiesti da Genesees V. 3.0 ed è riportata la procedura d'installazione e quella di avvio del software.*

2.1 Requisiti hardware e software e modalità di installazione

Genesees è un software sviluppato utilizzando il SAS SYSTEM V. 8.1 per Microsoft Windows, ovvero un package di uso generale che incorpora statistiche e procedure di analisi dei dati. Per utilizzare Genesees è necessario che sia installato il sistema SAS versione 8.1 ed in particolare i moduli: **SAS Language and Macro-facility, SAS IML Language, SAS STAT, SAS GRAPH.**

Lo spazio sul disco fisso necessario per l'installazione è di circa 4 MB ed è consigliabile una memoria di almeno 64 MB. Il tempo d'esecuzione della procedura è legato, ovviamente, alla velocità del processore installato e alla dimensione e complessità dei dati da elaborare.

L'utente riceve un CD-ROM di installazione corredato di un programma per installare il software.

Il software è disponibile anche effettuando il **download** (*aggiornamento 2005*):

- via internet (per utenti esterni all'istat):
<http://www.istat.it/Metodologi/index.htm> (selezionare “Metodi e Software per indagini statistiche”).
- via intranet (per utenti istat).

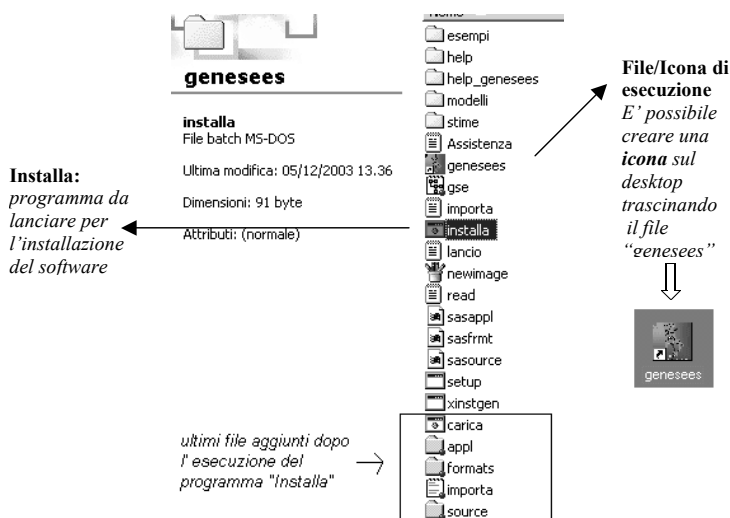
Propedeutica all'installazione del software è – ovviamente - quella del SAS v. 8.1.

Sia il CD-ROM che il download del software permettono di ottenere un file compresso. Per proseguire con l'installazione è perciò necessario avere a disposizione un programma per espandere il file `genesees3.zip` nella cartella `c:\genesees`.

Attenzione: Il file `genesees3.zip` deve espandersi solo nella cartella `c:\genesees`; non è possibile variare il nome della cartella di installazione. Inoltre è necessario installare il software su ogni postazione di lavoro con la procedura di seguito descritta e non è consentito copiare i file, senza effettuare la procedura d'installazione.

La procedura di installazione richiede la sola esecuzione del file **Installa.bat**, che crea nuovi file necessari all'esecuzione dei programmi. Al termine dell'esecuzione la cartella contiene i file mostrati in *figura 2.1*.

Figura 2.1: Il contenuto della cartella `c:\genesees` d'installazione - successivamente all'esecuzione del programma "installa.bat"



Dopo l'espansione, nella cartella c:\genesees sarà disponibile il file **read.me**, che contiene le istruzioni da eseguire per l'installazione ed il file **Assistenza.txt**, in cui leggere informazioni utili per ricevere assistenza sull'uso del software (cfr. *paragrafo 2.3*). Nella cartella c:\genesees\Esempi è memorizzato il *data-set* SAS che è alla base della applicazione descritta nella Sezione III del presente manuale d'uso.

2.2 La procedura di avvio e la password di esecuzione

L'esecuzione del programma installa.bat deve essere effettuata anche nel caso di installazioni successive alla prima.

Una volta installato il programma, la cartella c:\genesees contiene il **file di collegamento "genesees"**, che può essere spostato sul *desktop* per creare l'icona di lancio (cfr. *figura 2.1*). Il software si avvia perciò *clickando* due volte sul file di collegamento "genesees" oppure utilizzando l'icona creata sul *desktop*.

Attenzione: nel file collegamento (o nelle proprietà dell'icona) può essere necessario modificare i riferimenti al SAS.

Infatti, per default il *Collegamento* che è nelle *Proprietà* del file o dell'icona, ha la seguente *Destinazione* :

```
"C:\Programmi\SAS Institute\Sas\V8\sas.exe" -nologo  
-config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas.
```

Se l'utente – ad esempio – ha installato il SAS nel disco D, dovrà cambiare il percorso del file Sas.exe (attenzione: non quello del file lancio.sas o del file gse.cfg, che devono sempre essere riferiti alla cartella c:\genesees).

In dettaglio, il percorso aggiornato deve essere il seguente:

```
"D:\Programmi\SAS Institute\Sas\V8\sas.exe" -nologo  
-config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas.
```

La proprietà del file di collegamento o della icona sul desktop si varia utilizzando il bottone destro del mouse. Tra le voci che appaiono, selezionare "*Proprietà*" e poi "*Collegamento*", dove si legge, nel campo "*Destinazione*" il percorso di cui sopra.

Una volta installato, alla prima esecuzione Genesees chiede all'utente di contattare la struttura che si interessa dello sviluppo e della distribuzione del software generalizzato per ricevere una **password**.

In Istat, l'unità MTS/F si occupa dello sviluppo e distribuzione dei software generalizzati a supporto della produzione statistica nell'ambito del servizio Metodologie e Tecnologie e Software per la Produzione dell'Informazione Statistica (MTF).

Per garantire una tempestiva risposta alle esigenze dell'utenza (sia per ciò che concerne i **problemi tecnici** che per una veloce e controllata **diffusione di password e aggiornamenti**), l'unità ha messo a disposizione il seguente indirizzo di posta elettronica : mts-f@istat.it

Il software dunque mostra una maschera che riporta un **codice numerico** e richiede la password di registrazione. Per riceverla, è necessario contattare l'indirizzo di cui sopra, indicando il codice numerico.

Attenzione:

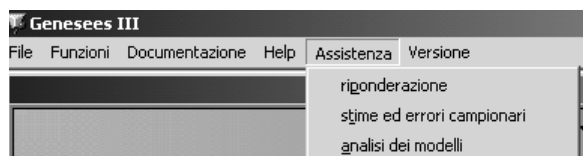
tale password è a servizio dell'utenza: in tal modo è possibile tenere traccia della lista degli utenti e, di conseguenza, inviare loro eventuali aggiornamenti del software.

Dopo la prima esecuzione, le successive installazioni per aggiornamento del software non richiederanno nuove password.

2.3 L'assistenza al software

Sia nel file Assistenza.txt che è nella cartella c:\genesees che nella voce *Assistenza* della schermata principale (cfr. figura 2.2) vengono riportate alcune informazioni utili all'utente che utilizza il software Genesees.

Figura 2.2: La voce Assistenza nel Menu di Genesees V. 3.0



Tali informazioni sono anche evidenziate nelle pagine intranet e internet agli stessi indirizzi riportati nel *paragrafo 2.1* che permettono di effettuare il download del software.

Le informazioni generali e quelle specifiche per la funzione di Stime ed Errori vengono di seguito riportate.

INFORMAZIONI SU PROBLEMATICHE RICORRENTI e CONTATTI:

Problemi relativi alla versione dei dataset di input

Assicurarsi di aver creato tutti i dataset di input in **versione SAS V8** (estensione SAS7BDAT oppure SD7). I dataset in versione SAS V6 (SD2) non sono gestiti correttamente dal software.

Problemi connessi al funzionamento del software

Errori durante l'installazione

1. Ad installazione terminata verificare il log di nome **Importa** presente nella cartella c:\genesees ed accertarsi che tutti i passi siano terminati con successo. In caso di errori segnalati nel log si consiglia di ripetere tutta la procedura di installazione (non è necessario disinstallare quanto già installato).
2. Se l'installazione è terminata con successo ma non si riesce ad avviare il software, è possibile che nel FILE DI COLLEGAMENTO (o nella icona di collegamento) sia necessario modificare i riferimenti al SAS.

Utilizzando il bottone di destra del mouse sul file di collegamento, è possibile andare in "Proprietà" e da qui in "Collegamento", dove si legge la "Destinazione":

"C:\Programmi\SAS Institute\Sas\V8\sas.exe" -nologo -config c:\genesees\gse.cfg -autoexec c:\genesees\lancio.sas".

Questo percorso deve essere modificato se il SAS è installato su disco o cartelle diverse da quelle di default (ad esempio se SAS è installato sul disco D).

FUNZIONE DI STIME ED ERRORI

Errori sulla creazione del dataset di input

Controllare sempre il dataset contenente le varie tipologie di errore intercettate e segnalate dal software, per avere informazioni sul tipo di errore commesso. In particolare, il software scrive il dataset ERRORI_INPUT nella cartella di output, ove memorizza gli errori rilevati sull'input.

File di log

Sia nel caso di avvertimento che di errore, uscendo dalla procedura si deve consultare il file genesees.log presente nella cartella di output, che contiene appunto il log della elaborazione effettuata.

Attenzione !

Se nel log appaiono messaggi del tipo:

NOTE: Invalid argument(s) to the exponential operator "***" at line 1554 column 14.

NOTE: Invalid argument(s) to the exponential operator "***" at line 1564 column 24.

SOTTOCLA=0 MODSCL=0 VARIABIL=totcosti MODALITA=1 _TYPE_=31 _FREQ_=3
OSSERVAZ=10 UP=10 uf=10 COMUNI=10 VARFIN=-1.828305E15
VARDIR=-1.881832E15 VARCLA=1.7332506E16

STIMA=209081036.77 TOTALE2=1.8115081E16 POP=18.021598705
POPCL=18.021598705 CAMPCL=10 DOMST=64 VARSRS=1.2585397E16

SQM=29505728.208 ERRAS=. ERRCL=131652974.3 ERREL=. ERRELPC=. LIMINF=.
LIMSUP=. DEFT=. EFFSTIM=0 B=1 RHO=0 _ERROR_=1 _N_=41

Missing values were generated as a result of performing an operation on missing values.

Each place is given by: (Number of times) at (Line):(Column). 4 at 1558:16 4 at 1559:15

4 at 1560:14 4 at 1560:20 4 at 1561:14 4 at 1561:20

Mathematical operations could not be performed at the following places. The results of the operations have been set to missing values.

Each place is given by: (Number of times) at (Line):(Column).

4 at 1554:14 4 at 1564:24

ciò dipende dal fatto che nel software per calcolare la varianza campionaria si utilizzano delle approssimazioni e - in casi eccezionali - può accadere che si ottengano valori negativi, da scartare.

CONTATTI in ISTAT

Per ricevere assistenza sul software GENESEES (aggiornamento 2005)

Per ERRORI imputabili al software *(non relativi alla creazione dei data-set di input) inviare una e-mail circostanziata, allegando i/il file di input, il dataset SAVEPAR e il log della elaborazione a:*

Roberto Di Giuseppe - Unità Software Generalizzati per la produzione statistica - MTS / F - digiusep@istat.it

È preferibile inviare una copia del messaggio anche all'indirizzo:

mts-f@istat.it

(Indirizzo Operativo dell' Unità Software Generalizzati per la produzione statistica - MTS / F)

Per problemi connessi con l'INSTALLAZIONE *utilizzare il seguente indirizzo:*

mts-f@istat.it (Indirizzo operativo dell' Unità Software Generalizzati per la Produzione Statistica - MTS / F)

Per problematiche metodologiche

Periodicamente vengono organizzati dei CORSI sugli aspetti metodologici e di utilizzo del software Genesees.

Per gli aspetti metodologici il responsabile è

Stefano Falorsi: stfalors@istat.it

Per avere informazioni che riguardano la CREAZIONE DEI DATA-SET DI INPUT ed in generale per PROBLEMI NON INFORMATICI i contatti consigliati sono:

Paolo Rigbi: parigbi@istat.it

Fabrizio Solari: solari@istat.it

Contatti in Istat: Informazioni generali

Nei precedenti punti è possibile identificare i giusti contatti da utilizzare per informazioni o problematiche riguardanti i software di interesse.

Se tali contatti non fossero quelli richiesti, per ricevere le adeguate indicazioni circa gli esperti informatici e metodologi da contattare, così come per informazioni generali sulle attività di sviluppo software generalizzati per la produzione statistica rivolgersi a:

Daniela Pagliuca - Responsabile Unità Operativa MTS/F “Software generalizzati per la produzione statistica”: pagliuca@istat.it

3. Il software Genesees V. 3.0: un insieme di funzioni

***Sintesi:** In questo capitolo viene illustrata la struttura del software Genesees V. 3.0 e vengono descritte le prime operazioni che l'utente deve attivare per avviare ed operare con il software*

3.1 La struttura del software Genesees V. 3.0

In questo capitolo viene illustrata la struttura del software Genesees V. 3.0, in modo tale che l'utente abbia una immediata visione del prodotto nel suo complesso. Viene descritta anche la schermata iniziale del software, tramite la quale è possibile selezionare le funzioni che lo compongono.

I prossimi capitoli sono invece dedicati alla specifica trattazione della funzione di Stime ed Errori, oggetto del presente manuale.

Propedeutica all'uso del software è - ovviamente - l'installazione (cfr. *capitolo 2*).

Genesees V. 3.0 viene attivato tramite il file "genesees" che si trova nella cartella c:\genesees d'installazione o tramite l'icona del programma che è stata creata sul desktop:

Figura 3.1: L'icona di avvio



Con l'avvio della procedura, si apre la schermata principale (cfr. *figura 3.2*), provvista di un menu, in cui compaiono le seguenti opzioni:

- **File:** per uscire dal software o richiamare una precedente elaborazione
- **Funzioni:** per attivare le funzioni principali del software
- **Documentazione:** per accedere alla documentazione on line, ovvero ai manuali di uso delle funzioni di Riponderazione, Stime ed errori campionari e Analisi dei Modelli.
- **Help** help-on-line sulla schermata di riferimento.
- **Assistenza:** prospetto riassuntivo dei problemi ricorrenti nell'utilizzo del software e contatti in Istat.
- **Versione** si riferisce all'ultima versione del software.

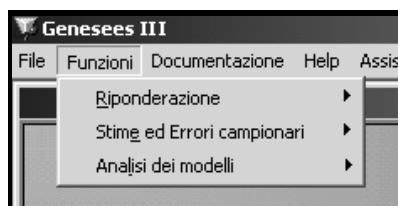
Figura 3.2 - La schermata principale



L'opzione **Funzioni** fornisce appunto la possibilità di accedere alle tre funzionalità principali implementate nel software (*cfr. figura 3.3*):

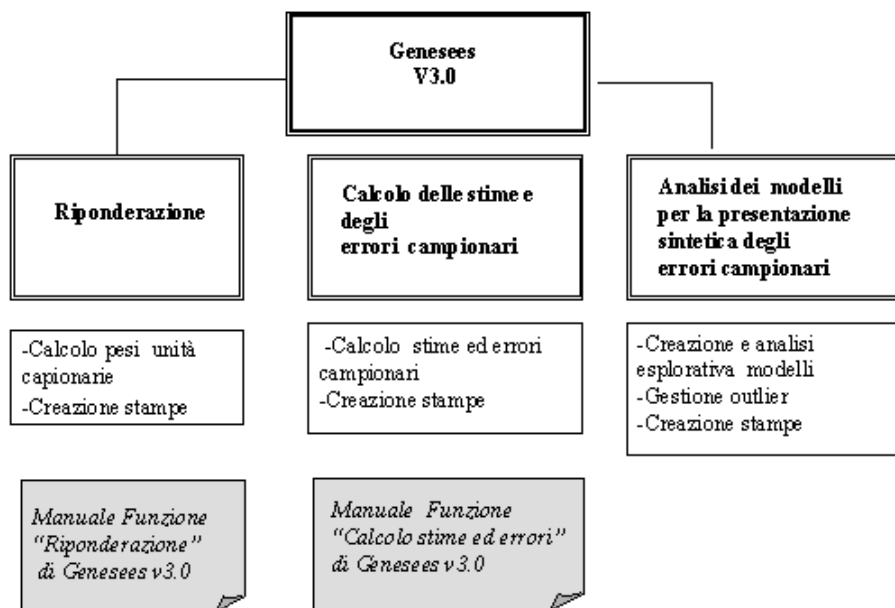
- Riponderazione
- Stime ed Errori campionari
- Analisi dei Modelli

Figura 3.3 - Le funzioni della schermata principale



Il software Genesees V. 3.0 è strutturato come mostrato nella successiva *figura 3.5*.

Figura 3.5 – La struttura del software Genesees V. 3.0



Come mostrato nella *figura 3.5*, la versione 3.0 comprende tre moduli. La funzione “Analisi dei Modelli” è l’ultimo modulo implementato in aggiunta alla versione 2.0 e tramite questa funzione viene dato ampio spazio alla presentazione grafica degli errori campionari.

Il presente manuale d’uso si riferisce esclusivamente alla funzione di Stime ed Errori attivata tramite l’opzione “Stime ed Errori campionari”

di figura 3.3. La funzione di “Riponderazione” è descritta in un manuale a sé stante (Pagliuca, 2004).

3.2 Le funzioni del software Genesees V. 3.0

La funzione di Riponderazione

La funzione di Riponderazione è applicabile in tutti i casi in cui esistono informazioni ausiliarie, espresse in termini di totali noti di variabili, definite appunto “ausiliarie”, legate alle variabili di interesse.

Essa è finalizzata al calcolo dei pesi finali da attribuire alle unità campionarie, sulla base di totali noti delle variabili ausiliarie e dei valori assunti da queste nel campione estratto.

Il contesto metodologico nel quale la funzione è stata concepita è quello degli stimatori di calibrazione (*calibration estimators*); tale teoria consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di calibrazione (Deville, J. C., Särndal, C. E., 1992, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, vol. 87, pp. 367-382).

La funzione di Calcolo Stime ed Errori

Lo scopo principale delle indagini campionarie è quello di fornire le stime di alcuni parametri descrittivi dell'intera popolazione, o di sottopopolazioni predefinite, dalla quale il campione viene estratto.

La funzione per il calcolo delle stime e degli errori campionari è finalizzata al calcolo delle stime e degli errori di campionamento e produce per ciascuna sottopopolazione di interesse: le stime oggetto di indagine e i corrispondenti errori di campionamento assoluti, relativi, e gli intervalli di confidenza; le principali statistiche che forniscono informazioni sull'efficienza della strategia di campionamento utilizzata (effetto del disegno ed effetto dello stimatore); i modelli di regressione per la presentazione sintetica degli errori di campionamento.

Anche tale funzione fa riferimento alla teoria degli stimatori di calibrazione (*calibration estimators*) e della relativa metodologia di calcolo della varian-

za; la metodologia consente di esprimere tutti gli stimatori utilizzati nelle indagini campionarie su larga scala, come casi particolari degli stimatori di calibrazione (Deville, J. C., Särndal, C. E., 1992, Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, vol. 87, pp. 367-382).

La funzione Analisi dei Modelli

La funzione di Analisi dei modelli nasce come estensione di quanto già implementato in Genesees v2.0 e aiuta l'utente a determinare la migliore rappresentazione sintetica degli errori campionari.

Tale funzione permette infatti di costruire i modelli per la presentazione sintetica degli errori di campionamento, come già era previsto nella versione 2.0 di Genesees, ma permette anche in aggiunta di analizzare la validità di tali modelli, in modo semplice ed interattivo.

La bontà di adattamento dei dati è facilmente migliorabile grazie al supporto di alcune funzionalità grafiche, che agevolano l'utente nel considerare alcuni valori come *estremi*, e grazie anche alla possibilità di procedere alla determinazione di un nuovo modello, che non tenga in considerazione i valori giudicati estremi, senza dover uscire dal software Genesees per modificare i dati di input eliminando i valori estremi.

4. La funzione Stime ed Errori campionari: cenni metodologici

La funzione di Stime ed Errori campionari di Genesee V. 3.0 fa riferimento alla metodologia costituita dalla classe degli *stimatori di calibrazione* (*calibration estimators*) e dalla relativa metodologia di calcolo della varianza (Deville e Särndal 1992).

Il software è pertanto in grado di rispondere, in modo metodologicamente valido, alla maggior parte dei problemi di stima e di calcolo degli errori campionari che si pongono nelle indagini campionarie effettuate dall'ISTAT sulle famiglie e sulle imprese.

In particolare il software produce i seguenti risultati:

- (a) calcolo di un insieme predefinito di stime d'interesse, a partire dai coefficienti finali di riporto presenti nel *data-set* di input ;
- (b) calcolo degli errori di campionamento assoluti, relativi percentuali e degli intervalli di confidenza per l'insieme prescelto di stime di cui al precedente punto (a);
- (c) costruzione dei modelli regressivi per la presentazione sintetica degli errori di campionamento, che legano gli errori campionari delle stime con i valori delle stime stesse;
- (d) calcolo dei valori interpolati degli errori campionari, in base ai modelli regressivi di cui al punto (c), per un insieme prefissato di valori tipici delle stime;
- (e) costruzione di alcune importanti statistiche utili sia per l'analisi critica della strategia di campionamento adottata, sia per la progetta-

zione di indagini future dello stesso tipo. Tra tali statistiche si ricordano: l'effetto del disegno di campionamento (*deff*) e l'efficienza dello stimatore adottato, indicato nel seguito come *effetto stimatore*.

Il sistema presenta una certa flessibilità nell'affrontare differenti problemi di stima della varianza in relazione a:

- *parametri d'interesse*, che possono essere totali e medie di variabili quantitative; frequenze assolute e proporzioni per le variabili qualitative; mediante opportuni accorgimenti nella formazione dell'input, è possibile considerare altri parametri quali rapporti e rapporti di rapporti;
- *disegni di campionamento*, che possono essere: di tipo casuale semplice; casuale a grappoli; casuale stratificato, semplice o a grappoli; a due o più stadi di selezione con eventuale stratificazione delle unità di primo stadio;
- *schemi probabilistici di selezione delle unità*, che possono essere con e senza reimmissione e con probabilità di selezione costante o variabile;
- *stimatori*, che rientrano nella classe generale degli stimatori di *calibrazione*;
- *domini di stima*, che rappresentano sottopopolazioni con riferimento alle quali sono fornite le stime e i corrispondenti errori di campionamento. Il software distingue i domini di stima in *pianificati* e *non pianificati* (o *sottoclassi*). I domini del primo tipo possono coincidere con gli strati del disegno, con aggregazioni di strati o con l'intera popolazione. I domini del secondo tipo sono sottopopolazioni che comprendono parzialmente le unità appartenenti agli strati del disegno.

Il software consente di trattare un'ampia classe di strategie di campionamento (disegno di campionamento e stimatore) per la stima di totali e medie di variabili quantitative e di frequenze assolute e relative di variabili qualitative, utilizzate nelle indagini su larga scala sia di tipo socio-demografico che di tipo economico.

La base metodologica a cui fa riferimento il software, per quanto riguarda gli stimatori, è quella relativa alla teoria degli *stimatori di calibrazione* (Deville e Särndal 1992; Singh e Mohl 1996). Tale base è del tutto generale perché tutti gli stimatori adottati nelle indagini ISTAT, e più in gene-

rale nelle indagini su larga scala condotte dai più importanti Istituti di statistica a livello internazionale, possono essere ottenuti come casi particolari, all'interno della famiglia degli *stimatori di calibrazione*. Vale la pena aggiungere, infatti, che possono essere ottenuti come casi particolari sia lo stimatore di *Horvitz-Thompson*, che è funzione lineare dei dati campionari, che tutti gli altri importanti stimatori non lineari che si utilizzano ogni qualvolta si disponga di informazioni ausiliarie esterne all'indagine, espresse sotto forma di totali noti. Tra gli altri si ricordano gli stimatori *rapporto*, *rapporto post-stratificato* e *regressione semplice*, che sono casi particolari della classe degli stimatori di *regressione generalizzata*, e gli stimatori *ratio raking* e *raking generalizzato* (nell'appendice A.1 sono contenuti i principali aspetti metodologici relativi a tali stimatori).

Tale classe di stimatori, in presenza di informazioni ausiliarie, permette di compensare generalmente gli errori di copertura, di mancata risposta totale e di migliorare l'efficienza delle stime.

Per quanto riguarda gli aspetti del disegno connessi con la stratificazione, il software tiene in considerazione, nella definizione dello stimatore della varianza, sia disegni di tipo semplice che stratificato (per approfondimenti si può consultare l'appendice A.3). Per i disegni a due o più stadi la stratificazione si riferisce alle unità di primo stadio e il calcolo della varianza non tiene conto di eventuali stratificazioni presenti negli stadi successivi. Nel caso in cui siano presenti strati con una sola unità campionaria, per calcolare la varianza viene adottata la tecnica del *collassamento degli strati* (Cochran 1977).

Stimatori lineari dei dati campionari

Considerando gli stimatori lineari - ed in particolare quello di *Horvitz-Thompson* (1952), che costituisce il metodo di stima di riferimento in assenza di informazioni ausiliarie - per un insieme di disegni campionari la procedura informatica impiega gli stimatori *corretti* della varianza campionaria più noti in letteratura (Cicchitelli et al. 1992, Särndal et al. 1992, Cochran 1977). Rientrano in questo insieme tutti i disegni a numerosità prefissata del seguente tipo:

- ad uno stadio, in cui si selezionano le unità con probabilità uguali e

quelli in cui le probabilità di selezione sono variabili e il processo di estrazione avviene con reimmissione (cfr. *tabella 4.1*);

- a due o più stadi, in cui si selezionano le unità primarie di campionamento con reimmissione e probabilità uguali o variabili (cfr. *tabella 4.1*).

Nel caso dei disegni ad uno stadio, in cui si selezionano le unità con probabilità variabili e senza reimmissione, e nel caso dei disegni a due o più stadi in cui le unità di primo stadio sono estratte senza reimmissione, il software utilizza lo stimatore della varianza campionaria relativo al caso di selezione delle unità di primo stadio con probabilità variabili e con reimmissione (cfr. *tabella 4.2*). Le principali giustificazioni per tale scelta operativa sono:

- l'elevata complessità di calcolo per ottenere una stima corretta della varianza campionaria, connessa principalmente con la determinazione delle probabilità di inclusione di secondo ordine;
- per alcuni metodi di selezione, inoltre, non è escluso che alcune probabilità di inclusione di secondo ordine si annullino determinando pertanto l'impossibilità di ottenere stimatori corretti della varianza (Brewer e Hanif 1982);
- per alcuni disegni a due o più stadi di selezione (che utilizzano ad esempio la selezione sistematica), spesso non esistono stimatori corretti delle componenti della varianza dovute agli stadi successivi al primo (Wolter 1985).

Applicando, in questo caso, la formula della varianza campionaria per disegni con reimmissione delle unità di primo stadio, non è richiesto il calcolo delle probabilità di inclusione del secondo ordine e la procedura di stima risulta, quindi, più rapida; inoltre, sebbene gli stimatori così definiti siano affetti da una distorsione positiva, questa risulta contenuta quando il tasso di campionamento è "piccolo". Infine, per i disegni a due o più stadi, tale approssimazione evita di calcolare esplicitamente le componenti di varianza dovute agli stadi successivi al primo (Wolter 1985).

Tabella 4.1: Disegni di campionamento (stratificati o non stratificati) in cui si adotta uno stimatore corretto della varianza campionaria per stimatori lineari o asintoticamente corretto per stimatori non lineari

Stadi di campionamento	Tipo di estrazione	Sistema probabilistico di selezione
Uno stadio con selezione di unità elementari (disegno casuale semplice)	con reimmissione	Probabilità uguali
	con reimmissione	Probabilità variabili
	senza reimmissione	Probabilità uguali
Uno stadio con selezione di grappoli di unità elementari (disegno casuale a grappoli)	con reimmissione	Probabilità uguali
	con reimmissione	Probabilità variabili
	senza reimmissione	Probabilità uguali
Due o più stadi	con reimmissione delle UPS [*] e con o senza reimmissione delle unità agli stadi successivi	Probabilità uguali delle UPS [*] e probabilità uguali o variabili delle unità agli stadi successivi
	con reimmissione delle UPS [*] e con o senza reimmissione delle unità agli stadi successivi	Probabilità variabili delle UPS [*] e probabilità uguali o variabili delle unità agli stadi successivi

*UPS: Unità di Primo Stadio

Tabella 4.2: Disegni di campionamento (stratificati o non stratificati) in cui si adotta uno stimatore approssimato della varianza campionaria

Stadi di campionamento	Tipo di estrazione	Sistema probabilistico di selezione
Uno stadio con selezione di unità elementari (disegno casuale semplice)	senza reimmissione	Probabilità variabili
Uno stadio con selezione di grappoli di unità elementari (disegno casuale a grappoli)	senza reimmissione	Probabilità variabili
Due o più stadi	senza reimmissione delle UPS [*] e con o senza reimmissione delle unità agli stadi successivi	Probabilità uguali delle UPS [*] e probabilità uguali o variabili delle unità agli stadi successivi
	senza reimmissione delle UPS [*] e con o senza reimmissione delle unità agli stadi successivi	Probabilità variabili delle UPS [*] e probabilità uguali o variabili delle unità agli stadi successivi

*UPS: Unità di Primo Stadio

Stimatori non lineari dei dati campionari

Il software consente di stimare la varianza di stimatori del totale, che ricorrono ad informazioni ausiliarie ed appartenenti alla famiglia degli stimatori di calibrazione. Per tali stimatori, che sono in generale funzioni non lineari delle osservazioni campionarie, non è nota l'espressione esatta dello stimatore corretto della varianza di campionamento.

Per ottenere un'espressione approssimata dello stimatore corretto della

varianza, il software impiega il *metodo della linearizzazione* proposto da Woodruff (Woodruff 1971, Cicchitelli et al. 1992) basato sull'espansione in serie di Taylor (Särndal et al. 1989, Deville e Särndal 1992). Il metodo, in sintesi, consiste nell'approssimare lo stimatore con una funzione lineare dei dati campionari (nell'*appendice A.2* sono contenuti i principali aspetti metodologici relativi al metodo della linearizzazione); effettuata tale trasformazione il software considera gli stimatori della varianza, già introdotti per lo stimatore di *Horvitz-Thompson* (cfr. *tabelle 1 e 2*).

Il metodo della linearizzazione è applicato facendo riferimento alla classe degli stimatori di regressione generalizzata, che si definiscono calibrando i coefficienti finali in base alla funzione di distanza euclidea. Per questi stimatori è possibile, infatti, ottenere un'espressione linearizzata.

Per quanto riguarda gli stimatori di calibrazione che fanno riferimento ad altre funzioni di distanza (ad esempio, *ratio raking*, *raking generalizzato*) – per i quali non è, invece, possibile ottenere la forma linearizzata dello stimatore – il software sfrutta la proprietà asintotica per cui tutti gli stimatori di calibrazione convergono alla classe degli stimatori di regressione generalizzata (Deville e Särndal 1992). Pertanto, la stima della varianza è calcolata considerando l'espressione linearizzata del corrispondente stimatore di regressione generalizzata in cui, tuttavia, i coefficienti finali sono quelli originati dal processo di calibrazione effettivo.

Principali statistiche prodotte

Il software fornisce le stime dei parametri di interesse e le rispettive varianze campionarie sia con riferimento ai domini pianificati che a quelli non pianificati (Cicchitelli et al. 1992). In particolare, per ciascuna stima l'output del software offre una serie di informazioni aggiuntive relative:

- al livello di precisione delle stime espresso in termini di errore assoluto di campionamento, errore relativo percentuale di campionamento, intervallo di confidenza al 95%;
- all'effetto del disegno di campionamento sulla precisione delle stime, espresso dalla statistica *deft* (Kish 1965) calcolata, per ciascuna stima di interesse, come radice quadrata del rapporto tra la stima della varianza della strategia adottata e la stima della varianza di un'i-

potetica strategia che prevede un campione casuale semplice di pari numerosità in termini di unità finali e lo stimatore espansione;

- all'efficienza dello stimatore utilizzato (effetto dello stimatore), espressa come radice quadrata del rapporto tra la stima della varianza ottenuta in base al disegno di campionamento ed allo stimatore adottato e la stima della varianza di una strategia che prevede il medesimo disegno campionario e lo stimatore espansione;
- alla correlazione intraclasse, valutata entro i grappoli per i disegni ad uno stadio a grappoli e valutata entro le unità primarie per i disegni a due o più stadi (Cicchitelli et al. 1992) .

Inoltre sono rese disponibili alcune indicazioni:

- sulla distribuzione della popolazione di riferimento, del campione osservato e delle unità primarie, per ciascun dominio di stima pianificato e per ciascuno strato;
- sulla formazione dei superstrati, ottenuti dal *collassamento* degli strati, nel caso in cui si sia resa necessaria l'adozione di tale tecnica per la stima della varianza (per approfondimento cfr. *paragrafo 5.1*).

Maggiori dettagli sulle statistiche fornite dal software sono illustrati nel *paragrafo 6*.

Infine, il software effettua una presentazione sintetica degli errori campionari stimati (cfr. *appendice A.5*), ottenuta secondo modelli regressivi che legano ciascuna stima al corrispondente errore di campionamento relativo o assoluto. Questa metodologia (Verma et al. 1980) è generalmente adottata nei volumi dell'ISTAT per documentare sinteticamente gli errori campionari delle stime ed evitare di pubblicare per ogni stima il corrispondente errore di campionamento relativo.

L'approccio utilizzato per la costruzione di questi modelli è diverso a seconda che si tratti di variabili qualitative o quantitative. Infatti, per quanto riguarda le stime di frequenze, è possibile utilizzare modelli che hanno un fondamento teorico secondo cui gli errori relativi delle stime di frequenze sono funzione decrescente dei valori delle stime stesse. Per quanto concerne le stime di totali di variabili quantitative, la definizione del modello interpolativo costituisce un problema di notevole complessità

perché non è stata elaborata una adeguata base teorica per l'interpolazione degli errori campionari delle stime in questione. L'approccio seguito nel software è pertanto di tipo empirico nel senso che si adattano diversi modelli regressivi che legano gli errori assoluti o relativi alle corrispondenti stime; tra i modelli stimati si sceglie quello che conduce ad un R^2 maggiore (Russo 1987).

Le principali caratteristiche metodologiche di tali prototipi sono contenute nei lavori seguenti: Falorsi e Falorsi (1995), Falorsi e Falorsi (1997), Falorsi e Falorsi (1998), Falorsi e Rinaldelli (1998), Falorsi, Pagliuca e Scepi (1999), Falorsi, Pagliuca e Scepi (2000), Pagliuca e Righi (2002), De Vitiis e Pagliuca (2003).

5. L'utilizzo della funzione di calcolo delle Stime e degli Errori di campionamento del software Genesees V. 3.0

Sintesi: Il capitolo 5 descrive in modo dettagliato l'utilizzo dell'interfaccia del software Genesees V.3.0 per il calcolo delle stime e degli errori di campionamento. I paragrafi 5.1 e 5.2 supportano l'utente nell'utilizzo delle maschere del software; il paragrafo 5.3 illustra la produzione delle stampe.

In particolare:

Il paragrafo 5.1 è introduttivo e spiega come avviare il software, riprendendo quanto già descritto nel capitolo 3 (riferito a Genesees V.3.0, visto nella sua globalità come insieme di funzioni).

Il paragrafo 5.2 entra nel merito della descrizione dell'uso della funzione di Stime ed Errori Campionari: il paragrafo 5.2.1 introduce le variabili e i parametri di input per la funzione di Stime ed Errori; nel paragrafo 5.2.2 e 5.2.3 è descritta la selezione di tali variabili di input.; il paragrafo 5.2.4 illustra come eseguire l'elaborazione vera e propria per ottenere le stime ed errori campionari.

Infine, nel paragrafo 5.3 sono riportate la selezione delle diverse tabelle e le informazioni che è possibile ottenere.

5.1 La schermata principale

Come premesso nel *capitolo 3*, il software Genesees V. 3.0 viene attivato tramite l'icona del programma posta sul desktop o tramite il file di collegamento "genesees", che si trova nella cartella c:\genesees d'installazione (per la procedura di installazione si consulti il *capitolo 2*).

L'avvio del programma mostra la schermata principale:

Figura 5.1 – La schermata principale

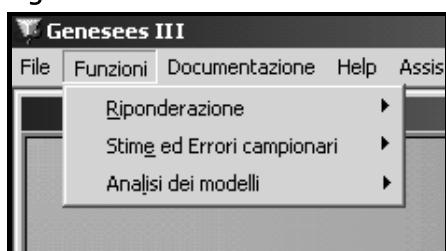


(M₀)

Tramite la voce **Funzioni** della schermata principale si possono attivare le tre funzioni di:

- Riponderazione
- Stime ed Errori campionari
- Analisi Modelli

Figura 5.2 – Le funzioni della schermata principale



In questo manuale viene trattata la **funzione Stime ed Errori campionari**, attivata tramite l'opzione omonima.

Come mostrato in *figura 5.3*, la funzione “Stime ed Errori campionari” permette a sua volta di attivare due opzioni:

- Calcolo Errori
- Creazioni Stampe

L'opzione “Calcolo errori” è utilizzata per il calcolo vero e proprio delle stime e degli errori di campionamento; “Creazione stampe” produce le stampe relative ad elaborazioni effettuate precedentemente.

Figura 5.3 - La funzione di calcolo delle stime e degli errori campionari



(M₀)

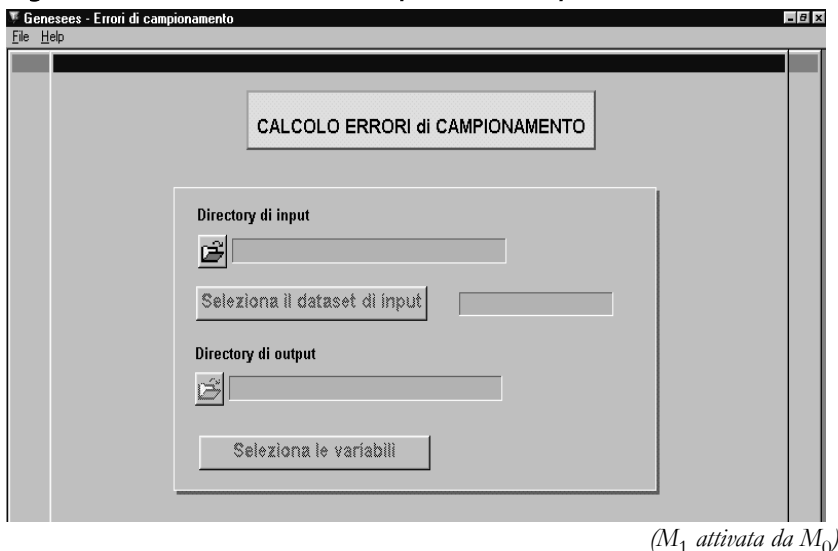
5.2 Il calcolo delle stime e degli errori campionari

L'opzione “Calcolo Errori” attiva la maschera M₁ di selezione dei parametri di input (cfr. *figura 5.4*).

Nella maschera M₁ è inoltre presente un menu bar con due voci:

- **File:** per uscire dal software
- **Help:** per visualizzare l'Help on line

Figura 5.4 - Maschera di selezione per i dati di input



Questa maschera consente di effettuare le seguenti scelte:

(1) Cartella e data-set di input



- scelta della cartella contenente il *data-set* SAS di input, utilizzando l'apposito bottone;
- selezione del *data-set* tra quelli contenuti nella cartella di input, utilizzando l'apposito bottone;

(2) Cartella di output



- si può scegliere anche la cartella di output che serve a memorizzare i *data-set* creati dalla procedura e gli eventuali file di stampa.

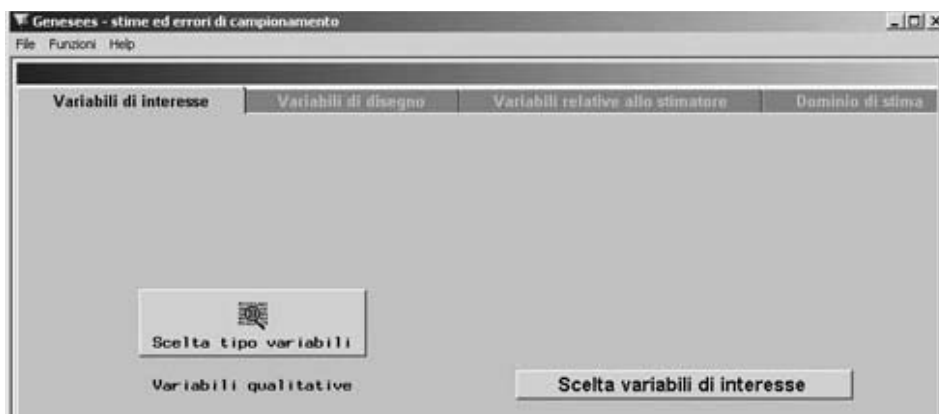
(3) Selezione delle variabili di input

Seleziona le variabili

- si possono infine selezionare le variabili dal *data-set* di input.

Questo ultimo bottone attiva la maschera M_2 (cfr. figura 5.5).

Figura 5.5 – Maschera di selezione delle variabili di input – Variabili di interesse

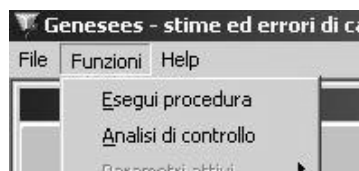


(M_2 attivata da M_1)

Nella maschera M_2 è presente un menu bar, in cui compaiono le seguenti voci:

- **File:** per tornare alla maschera precedente
- **Funzioni:** per eseguire la procedura, eseguire un'analisi di controllo dei dati di input, leggere i parametri di input
- **Help:** per visualizzare l'Help on line

Figura 5.6 - Opzioni della maschera M_2



In particolare la voce **Funzioni** comprende le tre opzioni (cfr. *figura 5.6*):

- Esegui Procedura
- Analisi di Controllo
- Parametri attivi

Le prime due voci “*Esegui procedura*” e “*Analisi di Controllo*” si attivano solo dopo la selezione delle variabili, da effettuare manualmente o in automatico; la terza voce “*Parametri attivi*” è utile nel caso in cui si siano già effettuate precedenti elaborazioni sugli stessi dati e si vogliano selezionare le medesime variabili in modo automatico.

La voce “*Esegui procedura*” è utilizzata per avviare il calcolo delle stime e degli errori di campionamento sulla base delle variabili selezionate nel *data-set* di input e creare i *data-set* di output.

La voce “*Analisi di controllo*” è utilizzata per effettuare una stampa a video di controllo dei dati di input. A tal proposito è da osservare che questa stampa è utile sia a priori, per verificare i dati di input prima dell’eventuale elaborazione (operazione che potrebbe presupporre tempi elaborativi piuttosto lunghi), che a posteriori; per questo è possibile creare la medesima stampa anche in una successiva fase, utilizzando la voce “*Creazione Stampe*” (cfr. *stampa 8, paragrafo 5.3*).

L’ultima voce “*Parametri attivi*” permette la selezione automatica delle variabili, ma solo se il *data-set* di input è stato precedentemente utilizzato e si voglia fare uso delle selezioni effettuate (cfr. *paragrafo 5.2.3*).

5.2.1 Le variabili e i parametri di input

Il funzionamento del software prevede la definizione di alcune variabili di input e di alcuni parametri. Le **variabili del *data-set* di input** possono essere raggruppate nelle seguenti tipologie, corrispondenti a quattro diverse schede nella maschera di selezione riportata in *figura 5.5*:

- A. Variabili di interesse**
- B. Variabili di disegno**
- C. Variabili relative allo stimatore**
- D. Variabili relative al dominio di stima**

Tramite le stesse schede si effettua anche la scelta di tre **parametri di input**:

- il tipo delle variabili di interesse: **qualitativo/quantitativo**;
- il **numero minimo di strati da aggregare** in un eventuale processo di *collassamento*;
- il **peso campionario**: *a livello di unità elementare o di cluster*.

La **costruzione del data-set di input** per il calcolo delle stime e degli errori campionari e la definizione delle variabili richieste, dipendono dal tipo di stimatore, dal disegno campionario utilizzato e dal livello di stima considerato. Essendo questa una operazione che esula dall'utilizzo vero e proprio delle maschere del software, si rimanda l'utente alla consultazione della Sezione II, avvertendolo che gli argomenti connessi con la costruzione del *data-set* di input implicano una conoscenza approfondita delle scelte metodologiche alla base del campione in esame.

Anche la **trattazione dei parametri di input** verrà approfondita nella successiva Sezione II.

Ai fini della successiva trattazione, si riportano di seguito le variabili di input da costruire e alcune informazioni sui parametri di input del software.

A. Variabili di interesse

- *Le variabili di interesse* – sono quelle per le quali si desiderano calcolare le stime e gli errori campionari

B. Variabili di disegno

- *Tipo di disegno* – codice relativo al tipo di disegno campionario
- *Unità primaria* – codice dell'unità primaria di campionamento
- *Unità finale* – codice dell'unità finale di campionamento nel caso di disegni a più stadi
- *Strato* – codice di strato nel caso di disegni stratificati
- *Peso diretto* – coefficiente diretto di riporto all'universo

C. Variabili relative allo stimatore

- *Peso distanza* – peso utile per definire lo stimatore
- *Variabili ausiliarie* utilizzate nello stimatore di calibrazione

- *Popolazioni pianificate utilizzate per lo stimatore* – codice identificativo delle partizioni di unità elementari che definiscono le popolazioni pianificate utilizzate per lo stimatore
- *Peso finale* – coefficiente finale di riporto all’universo

D. Variabili relative al dominio di stima

- *Variabili di sottoclassi* – variabili che definiscono i domini di stima non pianificati
- *Dominio pianificato* – codice relativo al livello di stima pianificato

5.2.2 La selezione delle variabili di input tramite la maschera di selezione

Le variabili obbligatorie devono essere tutte scelte per attivare le voci “Esegui procedura” e “Analisi di controllo”, presentate nel paragrafo precedente; in caso contrario saranno visualizzati opportuni messaggi d’errore.

La selezione manuale delle variabili è descritta in questo paragrafo.

• Variabili di interesse

La scheda relativa alla selezione delle “Variabili di interesse” è formata da due bottoni.

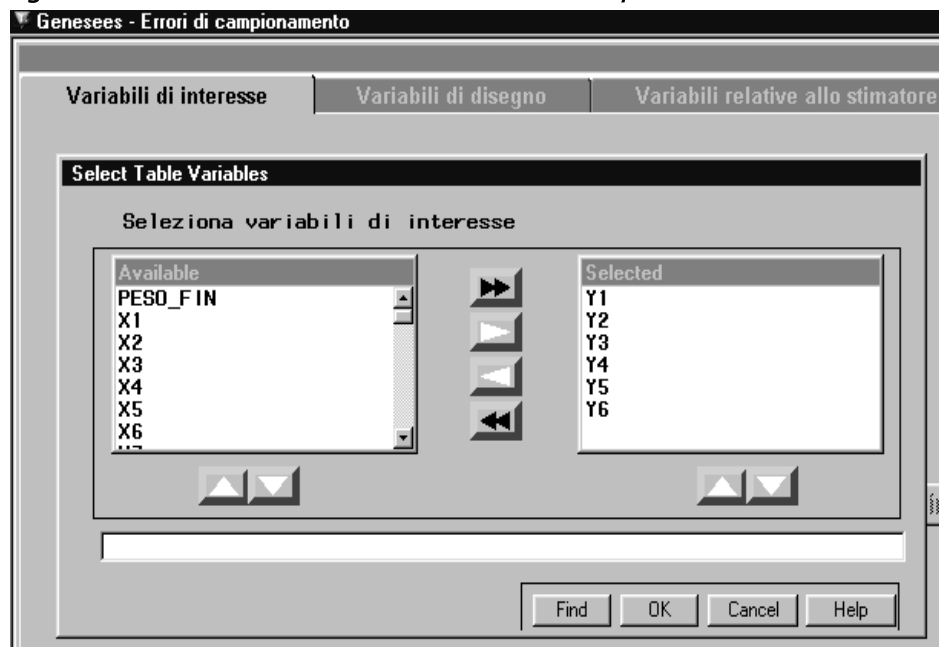


Un bottone deve essere utilizzato per specificare il **primo parametro di input**, ovvero se le variabili di interesse del *data-set* di input sono di tipo qualitativo o quantitativo (crf. *paragrafo 1.1.2, Sezione II*).




Il secondo bottone apre un’ulteriore maschera, visualizzata in *figura 5.7*, che mostra tutte le variabili di tipo numerico presenti nel *data-set* di input

Figura 5.7 – Maschera di selezione delle variabili di input – Variabili di interesse



(Maschera attivata da M_2)

 Le variabili possono essere selezionate singolarmente o in gruppo e spostate tramite la freccetta singola dal gruppo di sinistra (in cui vengono mostrate tutte le variabili disponibili – *available*) a quello di destra (in cui vengono poste le variabili selezionate – *selected*). I bottoni con le doppie freccette spostano tutte le variabili in entrambe le direzioni.

In figura 5.7 sono state selezionate alcune delle variabili di interesse Y.

Il pulsante “OK” conferma le operazioni effettuate, mentre il tasto “Find” consente di trovare una determinata variabile tra quelle presenti nel *data-set* di input.

- **Variabili di disegno**

Per inserire le variabili di disegno si deve attivare la seconda scheda:

Figura 5.8 – Maschera di selezione delle variabili di input – Variabili di disegno

Genesees - stime ed errori di campionamento

File Funzioni Help

Variabili di interesse Variabili di disegno Variabili relative allo stimatore Dominio

Numero minimo di unità per strato (unità di superstrato)

Tipo disegno

Unità primaria

Unità finale

Strato

Peso diretto

In questa scheda compare un **campo editabile** che presenta un valore di default pari a 2. Tramite questo è possibile variare il **secondo parametro** di input, ovvero il numero minimo delle unità che devono essere aggregate in un eventuale processo di *collassamento*. Il valore di default implica che il software, per formare il superstrato, aggrega tra loro coppie di strati originali; aggrega eccezionalmente tre strati, quando sono stati già formati superstrati ed è rimasto un singolo strato aggregabile; nella *Sezione II* (cfr. *paragrafo 1.1.2, Sezione II*) viene approfondito il concetto di collassamento e vengono specificate le condizioni che il software rispetta nell'aggregare due o più strati.

I cinque bottoni di questa scheda consentono di scegliere una sola variabile ognuno. La scelta è effettuabile analogamente a quanto descritto relativamente alle variabili di interesse.

A titolo esemplificativo viene riportata in *figura 5.9* la maschera di selezione della variabile “Tipo di disegno” (cfr. *paragrafo 5.1*) attivata tramite il corrispondente bottone

Tipo disegno

Figura 5.9 - Maschera di selezione della variabile "Tipo di Disegno"



- **Variabili relative allo stimatore**

La terza scheda serve ad inserire le variabili relative allo stimatore. La scelta è effettuabile analogamente a quanto descritto relativamente alle variabili di interesse.

Figura 5.10 – Maschera di selezione delle variabili di input – Variabili relative allo stimatore



(M_3 attivata da M_1)



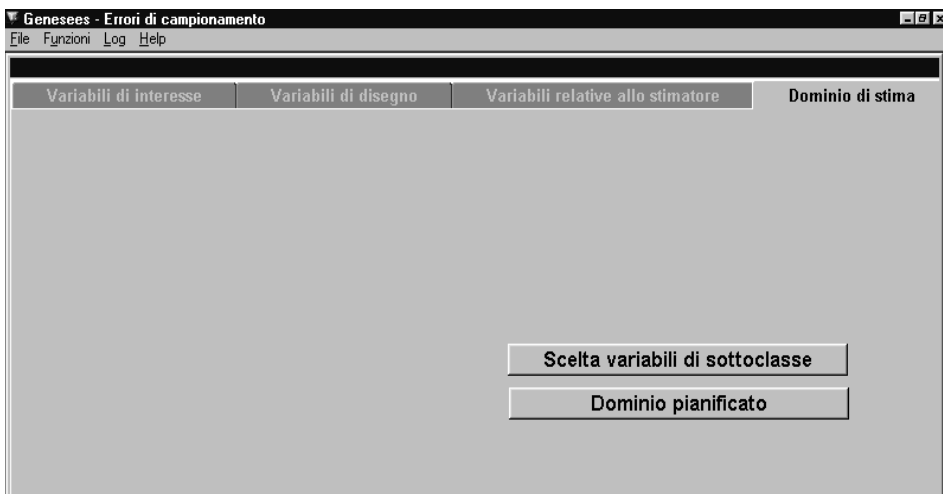
Tramite un bottone è possibile scegliere il **terzo parametro** di input, ovvero il tipo di peso; la procedura, infatti, consente di specificare se nel *data-set* di input sono stati considerati pesi *a livello di cluster* o pesi *a livello di unità elementare* (crf. *paragrafo 1.1.2, Sezione II*).

Gli altri bottoni aprono le rispettive maschere di selezione; fatta eccezione per la selezione delle variabili ausiliarie, negli altri casi è possibile effettuare una sola selezione.

- **Dominio di stima**

La quarta scheda (cfr. *figura 5.11*) concerne la selezione delle variabili relative al dominio di stima e presenta due bottoni da utilizzare per scegliere le “variabili di sottoclasse” (è consentito scegliere più variabili) o per scegliere la variabile corrispondente al “dominio pianificato” (si sceglierà un’unica variabile, cfr. *paragrafo 5.1*). La scelta è effettuabile analogamente a quanto descritto relativamente alle variabili di interesse.

Figura 5.11 – Maschera di selezione delle variabili di input – Variabili relative al dominio di stima



Attenzione: una volta selezionate le variabili in una qualsiasi delle quattro schede, se si utilizza la voce “Uscita” da File, le selezioni effettuate vengono perse.

5.2.3 La selezione delle variabili di input tramite i parametri attivati dal software

Come visto nel *paragrafo 5.2.1* le variabili possono essere selezionate tramite la relativa maschera M_I (si veda *figura 5.5*).

Figura 5.12. L'opzione "Parametri attivi"



Esiste un'alternativa per agevolare l'utente: l'opzione "Parametri attivi" (cfr. *figura 5.12*).

Tale voce è utilizzabile solo se è stata effettuata una precedente elaborazione con lo stesso *data-set* di input (in altri termini è già stata utilizzata l'opzione "Esegui procedura" o "Analisi di controllo"), in quanto il software crea nella cartella di output il *data-set* SAVEPAR.sas7bdat che memorizza i parametri della elaborazione (cfr. *paragrafo 2.1, Sezione II*).

Per usufruire di tale possibilità, nella elaborazione successiva occorre scegliere le stesse cartelle di input ed output della elaborazione precedente (ovviamente scegliendo la stessa cartella di output in diverse elaborazioni, il programma sovrascrive i *data-set* precedentemente memorizzati).

Come mostrato in *figura 5.12*, l'utente può scegliere le due voci "Mostra parametri" o "Accetta parametri".

La prima opzione permette la visualizzazione di una maschera simile a quella riportata in *figura 5.13*, tramite la quale l'utente può visualizzare i parametri.

Successivamente, per accettare i parametri visualizzati, l'utente deve scegliere la voce "Accetta parametri".

Ciò significa che le variabili vengono automaticamente selezionate. E' poi possibile modificare qualche scelta.

Figura 5.13: Parametri attivi

Parametri utilizzati per calcolo errori		
	descr	parametro
1	INPUT	C:\applicazioneLiberr.Esempio
2	Variabili di interesse	Y1 Y2 Y3 Y4 Y5 Y6
3	Peso distanza	PESO_DIST
4	Variabili ausiliarie	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14
5	Dominio pianificato	PROVINCIA
6	Peso finale	PESO_FIN
7	Sottoclassi	SEX
8	Tipo di disegno	TIPO_DISE
9	Unità primaria	UN_PRIM
10	Unità finale	COD_FAM
11	Strato	STRATO
12	Peso diretto	PESO_INIZ
13	Popolaz.pianif.util. per stimatore	REGIONE
14	Nr. unità superstrato	2
15	Variabili: 1 quant., 2 qual.	2
16	Stima: 1 unità elem., 2 cluster	1

(Maschera attivata da M₂)

5.2.4 L'elaborazione

Prima di elaborare i dati, il software esegue in automatico una serie di controlli per verificare che siano stati rispettati tutti i vincoli sulle variabili del *data-set* di input richiesti per l'utilizzo del software, vincoli presentati nel *paragrafo 1.2* della *Sezione II*.

Nel caso in cui tali vincoli non siano stati rispettati, la procedura invia un messaggio di errore, scrive nella libreria di output il relativo *data-set* contenente i dati di errore (*data-set* ERRORI_INPUT, cfr. *paragrafo 2.2*, *Sezione II*) e blocca l'elaborazione.

Dopo i controlli di input, prima dell'elaborazione, il software esegue in automatico un processo di *collassamento* su quegli strati che presentano un'unica unità per strato. Nel *paragrafo* precedente è possibile vedere il campo editabile che appare in *figura 5.8*, tramite il quale l'utente può variare il numero di strati da aggregare, numero che per default è pari a 2.

Come scritto nel precedente *paragrafo 5.2.2*, nel calcolare le stime e gli errori di campionamento, il software in automatico verifica se esistono strati con un'unica unità primaria; per tali strati non è infatti possibile calcolare la varianza e, di conseguenza, la stima della varianza riferita a quello strato corrisponderebbe ad un valore omesso, mentre la stima della varianza finale riferita ad un'eventuale partizione che includa tale strato risulterebbe sottostimata.

Il software prevede che per tali strati avvenga **in automatico** un processo di *collassamento*, formando alcuni “**superstrati**” (nella *Sezione II* - cfr. *paragrafo 1.1.2* – viene approfondito il concetto di collassamento e vengono specificate le condizioni che il software rispetta nell'aggregare due o più strati).

Dopo il processo di aggregazione il software effettua **un secondo controllo automatico**: verifica che tale aggregazione non sia fallita per alcuni strati che rimangono con un'unica unità primaria.

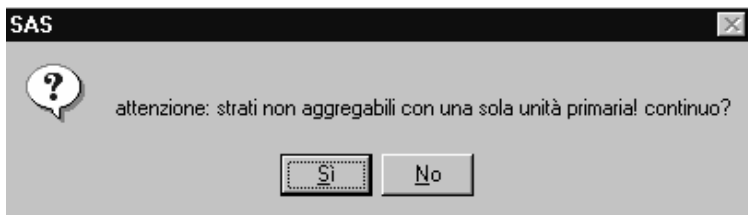
Il suddetto controllo è possibile sia tramite l'opzione “Esegui procedura” che tramite l'opzione “Analisi di controllo”. In tal caso il software mostra gli strati non aggregabili, per mezzo di una finestra come quella riportata di seguito:

Figura 5.14: Finestra in cui appaiono gli strati con una sola unità

Strati con una sola unità primaria				
	strato originale	pop.pianif.util. per stimato	dominio pianificat	peso diretto ▲
1	str02	REG1	PROV1	555
2	str07	REG2	PROV3	7760

La procedura **non si blocca** perché il software permette all'utente di scegliere se proseguire o meno, inviando un messaggio (cfr. *figura 5.15*). E' chiaro che, nel caso in cui l'utente decida di proseguire, deve considerare che la varianza risulterà sottostimata.

Figura 5.15: Messaggio che permette all'utente di fermare la procedura



Dopo questo controllo il software procede con l'elaborazione richiesta.

5.3 La funzione “Crea stampe”

La funzione di calcolo delle stime ed errori campionari viene attivata tramite la voce “Stime ed Errori campionari” della schermata principale. Essa consente di selezionare anche l'opzione “Crea stampe” per eseguire le stampe. Tale voce attiva a sua volta la maschera M_3 di figura 5.16. Le stampe ottenibili si riferiscono a dati elaborati precedentemente tramite la voce “Calcolo errori” (cfr. *paragrafo 5.2*). La cartella di input per le stampe corrisponde alla cartella che contiene i *data-set* di output di una precedente elaborazione.

Una volta selezionata la cartella, si scelgono le stampe desiderate tramite i bottoni dove appare il “SI”, valore che appare per *default* e che può essere variato. Se ad esempio si volesse ottenere solo la stampa numero 1 “Stime ed errori”, si dovrà fare in modo che appaia il “SI” sul bottone relativo alla prima stampa, mentre per le altre stampe si varia il valore di *default* da “SI” a “NO”.

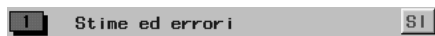
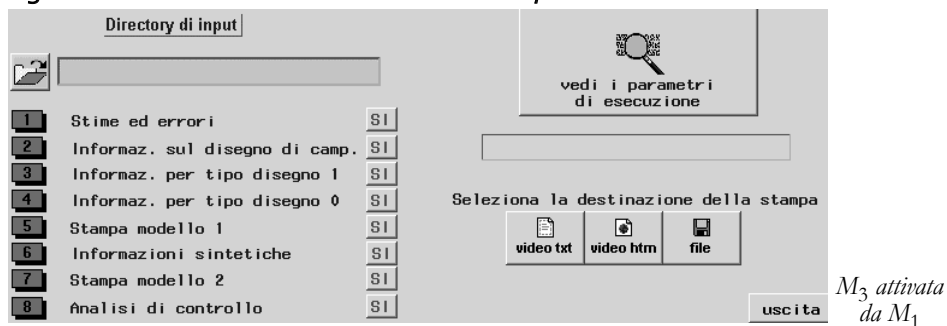


Figura 5.16 – Maschera di selezione delle stampe



Essendo possibile elaborare le stampe in un momento successivo a quello in cui sono stati creati i *data-set* di output, è sempre consentita la visualizzazione dei parametri di esecuzione della procedura.



Si può scegliere di effettuare, con appositi bottoni, le stampe a video¹ in formato txt o htm, oppure produrre delle stampe su file. In questo ultimo caso, vengono creati alcuni file ascii nella stessa cartella in cui sono stati memorizzati i *data-set* di output della precedente elaborazione.



Le stampe numero 5 “Stampa modello 1” e numero 7 “Stampa modello 2” inoltre creano in ogni caso quattro file excel utili per creare tabelle esterne al software (per approfondimenti cfr. *paragrafo 7*).

In questo paragrafo vengono elencate le informazioni che è possibile ottenere dalla funzione di calcolo delle stime ed errori campionari per fornire una documentazione da consultare velocemente; di seguito vengono anche riportate la prime schermate che appaiono per ciascuna delle stampe richieste a video.

Nel successivo capitolo 6 saranno approfondite le informazioni che è possibile ricavare tramite tali stampe.

La prima stampa - presentata in *figura 5.17* – riporta, per ciascuna variabile di interesse e con riferimento alle diverse modalità per le variabili qualitative:

1. il calcolo della stima
2. l'errore standard
3. l'errore relativo percentuale
4. i limiti dell'intervallo di confidenza (al livello di fiducia pari a 0,95).

Per approfondimenti cfr. *paragrafo 6.1 capitolo 6*.

Tali stime ed errori sono presentati con riferimento a ciascuno dei domi-

¹ Si noti che le stampe a video, una volta prodotte, vengono poste in secondo piano rispetto alla maschera attiva. Per portarle in primo piano, è sufficiente cliccare con il mouse in un qualsiasi punto nell'area di visibilità della stampa. Per abbandonare le stampe e tornare alla procedura si può usare il tasto PF3

ni pianificati per i quali si desidera ottenere la stime del totale delle variabili di interesse; se previsto, sono inoltre presentate anche per ciascuno dei domini non pianificati, ossia anche con riferimento a quelle partizioni della popolazione definite dalle modalità della variabile di sottoclasse (per approfondimenti sulla definizione dei domini di stima cfr. *capitolo 1, Sezione II*).

Figura 5.17: Stampa 1 - Stime ed errori di campionamento per dominio pianificato

Genesee - Errori di campionamento						
Results Viewer - SAS Output						
1 - Stime ed errori di campionamento per dominio di stima variabili qualitative						
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0						
variabili di interesse	modalità variabili interesse	stima	errore standard	errore relativo %	limite inf. I.C.	limite sup. I.C.
Y1	0	318150.70	3617.56	1.14	311060.3	325241.1
Y1	1	216604.10	3617.56	1.67	209513.7	223694.5
Y2	0	336971.00	3764.99	1.12	329591.6	344350.4
Y2	1	197783.80	3764.99	1.90	190404.4	205163.2
Y3	0	472925.40	3835.30	0.81	465408.2	480442.6
Y3	1	61829.40	3835.30	6.20	54312.20	69346.60

Come è possibile vedere dalle *figure 5.18, 5.19, 5.20*, le stampe 2, 3 e 4 riportano alcune informazioni riferite rispettivamente a tutte le unità - senza distinguere se sono estratte con diversi disegni campionari - (stampa 2), alle unità per le quali la variabile “Tipo di disegno” è pari ad 1 (stampa 3) e a quelle per le quali la variabile “Tipo di disegno” è pari a “0” (stampa 4).

La stampa 2 (*figura 5.18*) presenta, per ciascuna variabile di interesse e con riferimento alle diverse modalità per le variabili qualitative, alcune informazioni che hanno significato quando un disegno campionario è di tipo *composto* (per comprendere cosa si intenda in questa sede per disegno *com-*

posto si può leggere quanto scritto per la variabile “Tipo di disegno adottato” nel *paragrafo 1.1.1, Sezione II*):

5. lo scarto quadratico medio
6. il deft
7. l'effetto dello stimatore
8. il numero delle unità elementari
9. la stima del totale delle unità elementari

Le stampe 3 e 4 (5.19, 5.20) riportano, sempre per ciascuna variabile di interesse e con riferimento alle diverse modalità per le variabili qualitative, le informazioni presentate nella stampa2, con l'aggiunta di altre informazioni specifiche a seconda del disegno (stampa 3 “Tipo disegno”=1, stampa 4 “Tipo disegno”=0):

10. lo scarto quadratico medio
11. il deft
12. l'effetto dello stimatore
13. la correlazione intraclasse
14. il numero delle unità elementari
15. la stima del totale delle unità elementari
16. il numero di unità primarie
17. il numero medio di unità primarie

Per approfondimenti cfr. *paragrafi 6.3 e 6.4, capitolo 6*.

E' da osservare che le stampe 2, 3 e 4 riportano le suddette informazioni con riferimento a ciascuno dei domini pianificati per i quali si desidera ottenere la stima del totale delle variabili di interesse; se previsto, sono inoltre presentate anche per ciascuno dei domini non pianificati, ossia anche con riferimento a quelle partizioni della popolazione definite dalle modalità della variabile di sottoclasse (per approfondimenti sulla definizione dei domini di stima cfr. *capitolo 1, Sezione II*).

Figura 5.18: Stampa 2 - Informazioni sul disegno di campionamento per dominio di stima

Genesees - Errori di campionamento							
Results Viewer - SAS Output							
2 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative							
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0							
variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	numero unità elementari	stima del totale unità elementari	
Y1	0	0.491	0.76	0.17	3000	534755	
Y1	1	0.491	0.76	0.22	3000	534755	
Y2	0	0.483	0.80	0.17	3000	534755	
Y2	1	0.483	0.80	0.25	3000	534755	
Y3	0	0.320	1.23	0.12	3000	534755	
Y3	1	0.320	1.23	0.60	3000	534755	

Figura 5.19: Stampa 3 - Informazioni sul disegno di campionamento per dominio di stima – Tipo di disegno=1

Genesees - Errori di campionamento									
Results Viewer - SAS Output									
3 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative - tipo di disegno=1									
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0									
variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	correlaz. intraclasse	numero unità elementari	stima del totale unità elementari	numero di u. p.	n° medio per u.p.
Y1	0	0.491	0.94	0.58	-0.076	1849	262995	730	2.53
Y1	1	0.491	0.94	0.65	-0.076	1849	262995	730	2.53
Y2	0	0.484	0.96	0.57	-0.053	1849	262995	730	2.53
Y2	1	0.484	0.96	0.67	-0.053	1849	262995	730	2.53
Y3	0	0.310	1.52	0.49	0.863	1849	262995	730	2.53
Y3	1	0.310	1.52	0.96	0.863	1849	262995	730	2.53

Figura 5.20: Stampa 4 - Informazioni sul disegno di campionamento per dominio di stima – Tipo di disegno=0

Genesees - Errori di campionamento									
Results Viewer - SAS Output									
4 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative - tipo disegno=0									
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0									
variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	correlaz. intraclasse	numero unità elementari	stima del totale unità elementari	numero di u. p.	n° medio per u.p.
Y1	0	0.491	0.58	0.11	-0.008	1151	271760	13	88.54
Y1	1	0.491	0.58	0.14	-0.008	1151	271760	13	88.54
Y2	0	0.482	0.65	0.11	-0.007	1151	271760	13	88.54
Y2	1	0.482	0.65	0.17	-0.007	1151	271760	13	88.54
Y3	0	0.329	0.96	0.08	-0.001	1151	271760	13	88.54
Y3	1	0.329	0.96	0.45	-0.001	1151	271760	13	88.54

La stampa 5 produce i tabulati 5a e 5b mostrati in *figura 5.21 e 5.22*.

Per approfondimenti cfr. *paragrafi 6.5, capitolo 6*

La stampa 5a contiene, per ciascun dominio di stima pianificato, i coefficienti di regressione e di determinazione del modello utilizzato per la presentazione sintetica degli errori di campionamento per la stima di frequenze. Tale modello è descritto nella *appendice A.5*.

La stampa 5b riporta, distintamente per ciascun dominio di stima pianificato, gli errori relativi percentuali, interpolati secondo il modello e riferiti ad un insieme predefinito di valori tipici di stima. In altri termini sono presentati gli errori, calcolati secondo il modello, con riferimento a percentuali definite - percentuali che vanno dallo 0.1% fino al 50% della popolazione campionaria.

Figura 5.21: Stampa 5a - Valori dei parametri A e B ed indice di determinazione per dominio di stima pianificato del modello per la presentazione sintetica

Genesee - Errori di campionamento

Results Viewer - SAS Output

5a - Valori dei parametri A e B ed indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari

dominio pianificato	A	B	indice di determinazione
PROV1	7.2217	-1.59775	81.67
PROV2	11.0302	-1.64744	83.83
PROV3	11.3576	-1.70501	71.03
PROV4	10.9205	-1.68725	84.88
PROV5	11.9289	-1.76303	89.37
TOTALE	12.7455	-1.72878	88.30

Figura 5.22: Stampa 5b - Valori interpolati degli errori di campionamento per dominio di stima pianificato

Genesee - Errori di campionamento

Results Viewer - SAS Output

5b - Valori interpolati degli errori di campionamento per dominio di stima pianificato

	dominio pianificato										
	PROV1		PROV2		PROV3		PROV4		PROV5		TOTAL
	stima	errore rel.%	stima	errore rel.%	stima	errore rel.%	stima	errore rel.%	stima	errore rel.%	stima
stima %											
0.10	36.66	208.26	140.42	422.96	122.75	484.60	113.34	434.75	121.59	565.52	534.75
0.50	183.29	57.57	702.09	112.34	613.74	122.89	566.69	111.83	607.96	136.86	2673.77
1.00	366.58	33.09	1404.19	63.47	1227.48	68.06	1133.38	62.32	1215.92	74.29	5347.55
2.00	733.16	19.02	2808.37	35.86	2454.96	37.69	2266.76	34.73	2431.85	40.32	10695.10

La stampa 6 (figura 5.23) presenta le informazioni sintetiche sull'efficienza della strategia di campionamento adottata, che comprendono, distinta-

mente per ciascun dominio pianificato e per ognuna delle modalità delle variabili di sottoclasse prese in esame:

18. il deft medio
19. il deft massimo
20. l'effetto dello stimatore medio
21. l'effetto dello stimatore massimo
22. l'errore percentuale relativo medio
23. l'errore percentuale relativo massimo

Per approfondimenti cfr. *paragrafo 6.6, capitolo 6*.

Figura 5.23: Stampa 6 - Informazioni sintetiche sul campionamento per dominio di stima pianificato

Geneseees - Errori di campionamento						
Results Viewer - SAS Output						
6 - Informazioni sintetiche sul disegno di campionamento per dominio di stima						
dominio pianificato=TOTALE sottoclasse=0						
modalità sottoclasse	deft medio	deft massimo	effetto stim. medio	effetto stim. massimo	errore rel. % medio	errore rel. % massimo
0	1.14	1.45	0.32	0.90	3.5	16.3
dominio pianificato=TOTALE sottoclasse=SEX						
modalità sottoclasse	deft medio	deft massimo	effetto stim. medio	effetto stim. massimo	errore rel. % medio	errore rel. % massimo
1	0.79	1.41	0.36	0.97	4.3	20.3
2	0.81	1.52	0.43	0.92	5.3	18.1
dominio pianificato=PROV1 sottoclasse=0						
modalità	deft	deft	effetto	effetto	errore	errore

La stampa 7 produce i due tabulati 7a e 7b.

Per approfondimenti cfr. *paragrafo 6.6, capitolo 6*.

La stampa 7a (figura 5.24) contiene, per ciascun dominio pianificato, i coefficienti di regressione e di determinazione del modello alternativo utilizzato per la presentazione sintetica degli errori di campionamento per la stima di totali di variabili quantitative. Tale modello è descritto nella *appendice A.5*.

La stampa 7b (figura 5.25) riporta, distintamente per ciascun dominio di stima pianificato, gli errori relativi percentuali interpolati secondo il modello alternativo e riferiti ad un insieme predefinito di valori tipici di stima. In altri termini sono presentati gli errori, calcolati secondo il modello, con riferimento a percentuali definite - percentuali che vanno dallo 0.1% fino al 50% della popolazione campionaria.

Figura 5.24: Stampa 7a: Modello alternativo di interpolazione degli errori - Valori dei parametri ed indice di determinazione per dominio di stima pianificato

Geneseees - Errori di campionamento				
Results Viewer - SAS Output				
7a - Modello alternativo				
Valori dei parametri e indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari				
dominio pianificato	A	B	C	indice di determinazione
PROV1	180.91	0.011325	-.000000281	12.94
PROV2	1380.27	0.011084	-.000000064	6.54
PROV3	1293.89	0.008716	-.000000039	3.03
PROV4	955.54	0.012160	-.000000079	10.40
PROV5	1228.29	0.002742	0.000000018	8.97
TOTALE	2573.09	0.003556	-.000000003	7.74

Figura 5.25: Stampa 7b: Modello alternativo di interpolazione degli errori di campionamento per dominio di stima pianificato

Genesee - Errori di campionamento

Results Viewer - SAS Output

7b - Modello alternativo
Valori interpolati degli errori di campionamento per dominio di stima pianificato

	dominio pianificato										
	PROV1		PROV2		PROV3		PROV4		PROV5		
	stima	errore rel. %	stima	errore rel. %	stima	errore rel. %	stima	errore rel. %	stima	errore rel. %	sti
stima %											
0.01	3.67	4936.34	14.04	9830.82	12.27	10541.92	11.33	8432.10	12.16	10101.95	
0.02	7.33	2468.74	28.08	4915.96	24.55	5271.40	22.67	4216.66	24.32	5051.11	1
0.03	11.00	1646.20	42.13	3277.68	36.82	3514.55	34.00	2811.51	36.48	3367.50	1
0.04	14.66	1234.93	56.17	2458.53	49.10	2636.13	45.34	2108.94	48.64	2525.69	2

L'ultima stampa prodotta (*figura 5.26*) permette di avere informazioni circa l'eventuale processo di aggregazione degli strati. E' possibile ottenere questa stampa anche prima dell'elaborazione ("Analisi di controllo ", *paragrafo 5.2*).

La stampa presenta, per ciascuna popolazione pianificata utilizzata per lo stimatore e per ciascun dominio pianificato:

24. il tipo di aggregazione
25. il codice dello strato originale
26. il codice del superstrato, formato aggregando strati originali
27. il numero delle unità primarie
28. il numero delle unità finali
29. il numero delle unità elementari
30. il tipo di disegno
31. la stima del totale delle unità finali
32. la stima del totale delle unità elementari

Per approfondimenti cfr. *paragrafo 6.8, capitolo 6*.

Figura 5.26: Stampa 8 – Analisi di controllo sull’aggregazione degli strati: caso in cui i superstrati sono formati da 2 strati originari

Genesees - Errori di campionamento									
Results Viewer - SAS Output									
8 - Analisi di controllo sulla aggregazione degli strati: caso in cui i superstrati sono formati da 2 strati originari									
Popolaz. pianif. utiliz. per stimatore=REG1 dominio pianificato=PROV1									
tipo aggreg.	codice strato orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	tipo disegno	stima totale unità finali	stima totale unità elem.	
0	str01	1	399	399	964	1	14763	35668	
2	str02	2	1	15	36	0	555	1332	
dominio_pianificato				414	1000		15318	37000	
popolaz_pianificata				414	1000		15318	37000	
Popolaz. pianif. utiliz. per stimatore=REG2 dominio pianificato=PROV2									
tipo	codice	codice	numero	numero	numero	tipo	stima totale	stima totale	

6. La descrizione delle stampe

Al termine del procedura il software presenta i risultati ottenuti attraverso una serie di stampe che riportano per ciascuna stima di interesse, riferita ad un dato dominio di studio, la varianza ed alcune importanti statistiche utili per effettuare una analisi critica della strategia campionaria adottata.

Le espressioni matematiche delle statistiche presentate nei paragrafi seguenti utilizzano la stessa simbologia introdotta nelle appendici A.1 e A.3.

Avvertenze per una migliore lettura

Nei paragrafi che seguono vengono descritti tutti i campi delle stampe che sono mostrate nelle diverse figure del *paragrafo 5.3* (*cfr. figure 5.17-5.26, capitolo 5*); tali campi seguono la stessa numerazione progressiva, in modo da identificare facilmente la stampa a cui si riferiscono. E' inoltre da evidenziare che nel seguito – oltre che al suddetto *paragrafo 5.3* - si fa spesso riferimento anche all'*esempio 1.1 del paragrafo 1.3.1.1* e al *paragrafo 1.3.2* (e sottoparagrafi) della *Sezione II*; è dunque consigliabile prendere visione anche di tali paragrafi.

6.1 Stampa 1

Stime ed errori di campionamento per dominio di stima

Nella stampa sono presentate le stime dei totali e gli errori di campiona-

mento prodotti dal software per ciascuna variabile d'interesse e ciascun dominio di studio. In particolare, si hanno le seguenti statistiche:

1. **stima**: stima del totale della variabile, se questa è quantitativa, o stima della frequenza assoluta di ciascuna modalità della variabile, se questa è qualitativa;
2. **errore standard**: stima dello scarto quadratico medio dello stimatore;
3. **errore relativo %**: rapporto percentuale tra **errore standard** e **stima** relativi alla stessa variabile o, se questa è qualitativa, relativi alla stessa modalità della variabile;
4. **limite inf. I.C. e limite sup. I.C.**: limite inferiore e superiore dell'intervallo di confidenza al livello del 95% della stima;

Le espressioni matematiche delle statistiche elencate sono presentate nella *tabella 6.1*.

Tabella 6.1 - Descrizione delle statistiche della Stampa 1 per variabili di interesse di tipo quantitativo

Statistica	Simbolo	Formula di calcolo
Stima	A	$\sum_{k \in s_d} \frac{y_k}{\pi_k} g_k$
Errore standard	B	A seconda del disegno scelto è dato dalla radice quadrata di una tra le seguenti espressioni: (A.3.3), (A.3.5), (A.3.6), (A.3.8), (A.3.10), (A.3.12)
Errore relativo %	-	$(B/A) \times 100$
Limite inf. i. c.	-	$A - 1,96 \times B$
Limite sup. i. c.	-	$A + 1,96 \times B$

6.2 Stampa 2

Informazioni sul disegno di campionamento per dominio di stima

In questa seconda stampa si presentano per ogni variabile di interesse e dominio di studio alcune informazioni sulla strategia di campionamento adottata.

In particolare si hanno le seguenti statistiche:

5. **scarto q. medio**: stima dello scarto quadratico medio della relativa variabile a livello di dominio di stima considerato;
6. **deft**: radice quadrata del rapporto tra la stima della varianza della strategia adottata e la stima della varianza di una ipotetica strategia campionaria, che prevede un campione casuale semplice di pari numerosità, in termini di unità finali, al campione adottato e l'utilizzo dello stimatore espansione;
7. **effetto stimatore**: rapporto tra la stima della varianza per la strategia effettivamente utilizzata e la stima della varianza di una ipotetica strategia campionaria che prevede l'adozione del campione complesso utilizzato e lo stimatore espansione;
8. **numero di unità elementari**: numero di record presenti nel *dataset* di input relativo al dominio di stima considerato;
9. **stima del totale di unità elementari**: stima del numero di unità elementari appartenenti al dominio di studio considerato;

Le espressioni matematiche delle statistiche di cui sopra sono mostrate nella *tabella 6.2*.

Si avverte l'utente che nella *tabella 6.2* sono riportate anche altre informazioni - contrassegnate dal segno di asterisco – valide solo per le stampe 3 e 4, descritte nei prossimi due paragrafi.

Tabella 6.2 - Descrizione delle statistiche della Stampa 2, 3, e 4 per variabili di interesse di tipo quantitativo

Statistica	Simbolo	Formula di calcolo
Scarto q. medio	-	$\sqrt{\frac{\sum_{k \in s_d} \frac{y_k^2}{\pi_k} g_k}{\sum_{k \in s_d} \frac{g_k}{\pi_k}} - \left(\frac{\sum_{k \in s_d} \frac{y_k}{\pi_k} g_k}{\sum_{k \in s_d} \frac{g_k}{\pi_k}} \right)^2}$
Deft	C	$\sqrt{\frac{var(\hat{Y}_{GREG})}{var_{ccs}(\hat{Y}_{espansione})}}$ in cui $var_{ccs}(\cdot)$ è la stima della varianza dello stimatore nel disegno casuale semplice senza ripetizione.
Effetto stimatore	-	$\sqrt{\frac{var(\hat{Y}_{GREG})}{var(\hat{Y}_{espansione})}}$
Correlaz. Intraclassa*	-	$\frac{C^2 - 1}{G - 1}$ in cui G è il numero medio di unità elementari campione per unità primariati u. p..
Numero di unità elementari	E	Numero di record del data set
Stima del totale di unità elementari	-	$\sum_{k \in s_d} \frac{g_k}{\pi_k}$
Numero di u. p.*	F	Numero di unità primarie o di grappoli
N° medio per u. p.*	G	E/F

*Statistiche presenti nelle stampe 3 e 4.

6.3 Stampa 3

Informazioni sul disegno di campionamento per dominio di stima in cui la variabile “tipo di disegno” è pari a “1”

Le informazioni raccolte in questa stampe si riferiscono ai campioni selezionati secondo i disegni illustrati nei *paragrafi 1.3.2.3 e 1.3.2.4 della Sezione*

II, per i quali è opportuno porre la variabile “Tipo di disegno” pari a “1”. Alcune delle statistiche presenti coincidono nel significato con quelle introdotte nella stampa 2, riferendosi, tuttavia, agli strati di un dato dominio di stima in cui la variabile “Tipo di disegno” è pari a “1”.

In particolare, con riferimento al generico dominio di stima, si ha che:

- lo **scarto q. medio**, il **deft**, l'**effetto stimatore**, il **numero di unità elementari** e la **stima del totale di unità elementari** sono calcolati come nella stampa 2 (cfr. *paragrafo 6.2*);
- la **correlaz. intraclasse**: è la stima della correlazione media tra le unità elementari all'interno dei grappoli, per i disegni ad uno stadio a grappoli, o all'interno delle unità primarie, per i disegni a due o più stadi. La statistica viene anche detta coefficiente di omogeneità intraclasse.
- il **numero di u. p.**: poiché con “Tipo di disegno” pari a “1” si sta considerando un disegno ad uno stadio, la statistica indica il numero di grappoli se il disegno è a grappoli, oppure il numero di unità elementari per i disegni non a grappoli. In entrambi i casi la statistica fornisce il numero di unità finali;
- il **numero medio per u. p.**: è il numero medio di unità elementari appartenenti ad un grappolo. Naturalmente per i disegni non a grappoli la statistica deve essere pari a “1”.

Le espressioni matematiche delle statistiche sono illustrate nella *tabella 6.2*.

E' necessario ricordare che le statistiche che sono presenti nelle stampe si basano su tutti i record con la variabile “Tipo di disegno” adottato pari ad “1”. Pertanto se nel *data-set* sono presenti record selezionati con disegni differenti, ma la variabile “Tipo di disegno” adottato presenta lo stesso valore, le statistiche che si ottengono potrebbero essere fuorvianti.

6.4 Stampa 4

Informazioni sul disegno di campionamento per dominio di stima in cui la variabile “tipo di disegno” è pari a “0”

Le informazioni raccolte in questa stampa si riferiscono ai campioni sele-

zionati secondo i disegni illustrati nei *paragrafi* 1.3.2.1, 1.3.2.2 1.3.2.5, 1.3.2.6 e 1.3.2.7 nella *Sezione II*, per i quali è opportuno porre la variabile “Tipo di disegno” pari a “0”.

Alcune delle statistiche presenti coincidono nel significato con quelle introdotte nella stampa 2, riferendosi, tuttavia, agli strati di un dato dominio di stima in cui la variabile “Tipo di disegno” è pari a “0”.

In particolare, con riferimento al generico dominio di stima, si ha che:

- lo **scarto q. medio**, il **deft**, l'**effetto stimatore**, il **numero di unità elementari** e la **stima del totale di unità elementari** sono calcolati come nella stampa 2 (cfr. *paragrafo* 6.2);
- la **correlaz. intraclasse**: è la stima della correlazione media tra le unità elementari all'interno dei grappoli, per i disegni ad uno stadio a grappoli, o all'interno delle unità primarie, per i disegni a due o più stadi. La statistica viene anche detta coefficiente di omogeneità intraclasse.
- il **numero di u. p.**: poiché con “Tipo di disegno” pari a “0” si possono considerare disegni con diversi stadi di selezione, la statistica indica il numero di grappoli se il disegno è ad uno stadio a grappoli, oppure è il numero di unità primarie nei disegni a due o più stadi. Nel caso dei disegni ad uno stadio non a grappoli la statistica indica il numero di unità elementari.;
- il **numero medio per u. p.**: è il numero medio di unità elementari appartenenti ad un grappolo se il disegno è ad uno stadio. La statistica rappresenta il numero di unità elementari appartenenti all'unità primaria per i disegni a due o più stadi. Per i disegni ad uno stadio non a grappoli la statistica deve essere pari a “1”.

Le espressioni matematiche delle statistiche sono analoghe a quelle contenute nella *tabella* 6.2.

Particolare attenzione al significato delle statistiche deve essere posta quando le unità del *data-set* con “Tipo di disegno” pari a “0” sono state selezionate con diversi disegni di campionamento.

6.5 Stampa 5

Modelli interpolanti per la rappresentazione sintetica degli errori campionari per la stima di frequenze

- Stampa 5a: I valori dei parametri A e B e indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari

In questa stampa sono presentati i risultati della rappresentazione sintetica degli errori campionari, ottenuta mediante il metodo dei modelli regressivi (cfr. *appendice A.5*).

Nella tabella vengono riportati i valori dei parametri stimati e l'indice di determinazione del modello:

$$\log(\hat{\epsilon}^2(\hat{Y})) = \hat{\alpha}_1 + \hat{\alpha}_2 \log(\hat{Y})$$

che viene utilizzato nel caso di stime di frequenze.

I parametri A e B della stampa 5a corrispondono ai parametri e del modello di cui sopra.

In particolare si hanno le seguenti variabili:

- **dominio pianificato**: codici identificativi dei domini pianificati, su ciascuno dei quali è costruito il modello regressivo. L'ultima modalità "TOTALE" è relativa all'intera popolazione di riferimento;
 - **A** : stima ottenuta in base al metodo dei minimi quadrati del parametro del modello, con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
 - **B**: stima ottenuta in base al metodo dei minimi quadrati del parametro del modello con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
 - **indice di determinazione**: indice R^2 % del modello con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
- Stampa 5b - Valori interpolati degli errori di campionamento per dominio di stima pianificato

In questa seconda stampa, sempre relativa al modello regressivo per la rappresentazione sintetica degli errori delle stime di frequenze, si presentano gli errori relativi per alcune stime di frequenze assolute prefissate nei diversi domini pianificati. In particolare ciascuna delle stime di frequenze prefissate è una frazione della stima della popolazione delle unità finali calcolata con i coefficienti finali. Le variabili che compaiono nella tabella sono le seguenti:

- **stima %**: la variabile indica la frazione della stima del totale della popolazione delle unità finali calcolata con i coefficienti finali sulla quale sono forniti gli errori relativi. Le frazioni percentuali prese in considerazione sono: 0,1%, 0,5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%;
- **dominio pianificato**: codici identificativi dei domini pianificati, su ciascuno dei quali è costruito il modello regressivo. Le modalità della variabile sono presentate nella riga sottostante. L'ultima modalità "TOTALE" è relativa all'intera popolazione di riferimento;
- **stima**: frazione della stima della popolazione delle unità finali nel dominio pianificato identificato dal codice della variabile **dominio pianificato**. La frazione della stima è indicata dal valore della variabile **stima %**;
- **err. rel.%**: errore relativo percentuale della frequenza assoluta indicata dalla variabile **stima** per il dominio pianificato identificato dal codice della variabile **dominio pianificato**.

6.6 Stampa 6

Informazioni sintetiche sul disegno di campionamento per dominio di stima

Le tabelle presentate in questa stampa offrono alcune informazioni generali di sintesi sulla precisione delle stime prodotte con la strategia campionaria adottata dall'utente per ciascun dominio di studio.

- **deft medio**: media dei deft (stimati), calcolata considerando i deft di tutte le stime di interesse;

- **deft massimo:** deft (stimato) massimo ottenuto considerando i deft di tutte le stime di interesse;
- **effetto stim. medio:** media degli effetti stimatori, calcolata considerando gli effetti stimatori di tutte le stime di interesse;
- **effetto stim. massimo:** effetto stimatore massimo, ottenuto considerando gli effetti stimatori di tutte le stime di interesse;
- **errore rel. % medio:** media degli errori relativi percentuali, calcolata considerando gli errori relativi di tutte le stime di interesse;
- **errore rel. % massimo:** errore relativo percentuale massimo ottenuto considerando gli errori relativi di tutte le stime di interesse.

6.7 Stampa 7

Modelli interpolanti per la stima di totali di variabili quantitative

- Stampa 7a - Modello alternativo - Valori dei parametri e indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari

In questa stampa sono presentati i risultati della rappresentazione sintetica degli errori campionari, per stime di totali di variabili quantitative, ottenuta mediante il metodo dei modelli regressivi (cfr. *appendice A.5*). Nella *tabella 7a* vengono riportati i valori dei parametri stimati e l'indice di determinazione del modello:

$$\hat{\hat{e}}(\hat{Y}) = \hat{\alpha}_2 + \frac{\hat{\alpha}_1}{\hat{Y}} + \hat{\alpha}_3 \hat{Y}.$$

I parametri A, B e C della stampa corrispondono ai parametri , e del modello di cui sopra.

La tabella contiene le seguenti variabili:

- **dominio pianificato:** codici identificativi dei domini pianificati, su ciascuno dei quali è costruito il modello regressivo. L'ultima modalità "TOTALE" è relativa all'intera popolazione di riferimento;
- **A :** stima ottenuta in base al metodo dei minimi quadrati del parametro del modello, con riferimento al dominio pianificato identi-

cato dal codice della variabile **dominio pianificato**;

- **B:** stima ottenuta in base al metodo dei minimi quadrati del parametro del modello, con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
 - **C:** stima ottenuta in base al metodo dei minimi quadrati del parametro del modello, con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
 - **indice di determinazione:** indice R^2 % del modello, con riferimento al dominio pianificato identificato dal codice della variabile **dominio pianificato**;
- Stampa 7b - Modello alternativo - Valori interpolati degli errori di campionamento per dominio di stima pianificato

In questa seconda stampa, sempre relativa al modello regressivo per la rappresentazione sintetica degli errori delle stime di totali di variabili quantitative, si presentano gli errori relativi per le stime di totali prefissati nei diversi domini pianificati. In particolare, per ciascun dominio pianificato si considera la stima più elevata tra quelle calcolate per le diverse variabili di interesse. Della stima prescelta per ciascun dominio pianificato, si considerano diversi valori ottenuti come frazioni della stima stessa. Le variabili che compaiono nella tabella sono le seguenti:

- **stima %:** la variabile indica la frazione della stima del totale sulla quale sono forniti gli errori relativi. Le frazioni percentuali prese in considerazione sono: 0,01%, 0,02%, 0,03%, 0,04%, 0,05%, 0,1%, 0,5%, 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%;
- **dominio pianificato:** codici identificativi dei domini pianificati, su ciascuno dei quali è costruito il modello regressivo. Le modalità della variabile sono presentate nella riga sottostante. L'ultima modalità "TOTALE" è relativa all'intera popolazione di riferimento;
- **stima:** frazione della stima del totale nel dominio pianificato identificato dal codice della variabile **dominio pianificato**. La frazione della stima è indicata dal valore della variabile **stima %**;
- **err. rel. %:** errore relativo percentuale del totale, indicato dalla varia-

bile **stima** per il dominio pianificato identificato dal codice della variabile **dominio pianificato**.

6.8 Stampa 8

Analisi di controllo sulla aggregazione degli strati, caso in cui i superstrati sono formati da due strati originari

Nella stampa sono presentate una serie di tabelle relative al processo di aggregazione (*collassamento*) degli strati per poter ottenere la stima della varianza campionaria. Ciascuna tabella è relativa ad una delle combinazioni esistenti tra le modalità della variabile “Popolazione pianificata utilizzata per lo stimatore” (**Popolaz. pianif. utiliz. per stimatore**) e “Dominio pianificato”. Il processo di aggregazione degli strati può essere infatti effettuato solo tra gli strati che appartengono contemporaneamente alla stessa popolazione pianificata utilizzata per lo stimatore e allo stesso dominio pianificato (per approfondimento cfr. *paragrafo 1.1.2, Sezione II*).

Per ogni tabella si hanno le variabili seguenti:

- **tipo aggreg.**: il tipo di aggregazione che ha subito lo strato identificato dalla variabile **codice strato originale**. La variabile è pari a “0” per gli strati che non devono essere collassati; pari a “1” per gli strati che sono stati collassati; pari a “2” per gli strati che devono essere collassati ma che non è stato possibile collassare;
- **codice strato originale**: codice dello strato del disegno di campionamento. Gli strati elencati in ciascuna tabella sono quelli che appartengono alla particolare combinazione delle modalità della variabile “Popolazione pianificata” utilizzata per lo stimatore (**Popolaz. pianif. utiliz. per stimatore**) e “Dominio pianificato”;
- **codice superstr. finale**: codice assegnato dal software agli strati dopo che è stato effettuato il processo di aggregazione degli strati;
- **numero unità primarie**: numero di unità primarie nel campione (se il disegno è a due o più stadi) o numero di grappoli di unità (se il disegno è a grappoli) appartenenti allo strato identificato dalla variabile **codice strato originale**;

- **numero unità finali:** numero di unità finali nel campione relativo allo strato identificato dalla variabile **codice strato originale**;
- **numero di unità elem.:** numero di unità elementari (record presenti nel *data-set* di input) relativo allo strato identificato dalla variabile **codice strato originale**;
- **tipo disegno:** codice del disegno di campionamento nel quale è inserito lo strato identificato dalla variabile **codice strato originale**. Se il codice della variabile è pari a “1” lo strato fa parte di un disegno di campionamento stratificato senza reimmissione e con probabilità di inclusione nel campione costante. Se il codice è pari a “0” lo strato fa parte degli altri disegni di campionamento stratificati implementati nel software;
- **stima totale unità finali:** stima della popolazione delle unità finali relativa allo strato identificato dalla variabile **codice strato originale**;
- **stima del totale di unità elementari:** stima della popolazione delle unità elementari (identificate dai record del *data-set*) relativa allo strato identificato dalla variabile **codice strato originale**.

Oltre a queste informazioni le tabelle presentano altre informazioni campionarie per “Dominio pianificato” e “Popolazione pianificata utilizzata per lo stimatore”. In particolare in ciascuna tabella è presente una riga di totale identificata con:

- **dominio_pianificato:** in cui si trovano i totali delle variabili **numero unità finali**, **numero di unità elem.**, **stima totale unità finali** e **stima del totale di unità elementari**, degli strati appartenenti al “Dominio pianificato” a cui si riferisce la tabella (cfr. *figura 6.1*).

Figura 6.1 – Stampa 8 (1)

tipo aggreg.	codice strato orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	tipo disegno	stima totale unità finali
0	str08	4	2	86	238	0	21670
0	str09	5	2	47	120	0	12292
2	str07	6	1	40	142	0	7760
dominio_pianificato				173	500		41722

Per le tabelle relative all'ultimo "Dominio pianificato" contenuto in una data "Popolazione pianificata utilizzata per lo stimatore", si aggiunge alla riga **dominio_pianificato** una seconda riga (cfr. *figura 6.2*) indicata con:

- **popolaz_pianificata**: in cui si trovano i totali delle variabili **numero unità finali**, **numero di unità elem.**, **stima totale unità finali** e **stima del totale di unità elementari**, degli strati appartenenti alla "Popolazione pianificata utilizzata per lo stimatore". Tali totali sono ottenuti sommando i valori assunti da tali variabili nelle diverse tabelle che presentano la stessa modalità della variabile **Popolaz. pianif. utiliz. per stimatore**.

Figura 6.2 – Stampa 8 (2)

tipo aggreg.	codice strato orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	tipo disegno	stima totale unità finali	stima del totale di unità elementari
0	str01	1	399	399	964	1	14763	35
2	str02	2	1	15	36	0	555	1
dominio_pianificato				414	1000		15318	37
popolaz_pianificata				414	1000		15318	37

Infine l'ultima tabella (*figura 6.3*) della stampa oltre alle due righe indicate con **dominio_pianificato** e **popolaz_pianificata**, ne presenta una terza (senza intestazione) in cui sono calcolati, su tutti gli strati dell'universo, i totali delle variabili **numero unità finali**, **numero di unità elem.**, **stima totale unità finali** e **stima del totale di unità elementari**.

Figura 6.3 – Stampa 8 (3)

tipo aggreg.	codice strato orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	tipo disegno	stima totale unità finali	stima del totale di unità elementari
0	str14	10	80	80	244	1	18400	
1	str15	11	1	36	100	0	9360	
1	str16	11	1	36	105	0	9180	
1	str17	11	1	19	51	0	5168	
dominio_pianificato				171	500		42108	1
popolaz_pianificata				711	2000		180283	4
				1125	3000		195601	5

7. I file di output della funzione di Stime ed Errori di Genesees

Il software produce alcuni data-set di output e alcuni file ascii, scritti sulla cartella di output scelta dall'utente, alcuni file excel e produce infine il file "genesees.log".

Genesees produce otto **file ascii** che contengono le tabelle prodotte dal software e che - come specificato nel capitolo precedente - è possibile memorizzare in file esterni. I file sono i seguenti:

- stampa1.txt,
- stampa2.txt,
- stampa3.txt,
- stampa4.txt,
- stampa5.txt,
- stampa6.txt,
- stampa7.txt,
- stampa8.txt.

Tali file vengono ovviamente scritti **solo a richiesta** dell'utente, solo se ha selezionato la stampa corrispondente e utilizza il bottone "file".

A titolo di esempio, nella *Sezione III* - relativa alla applicazione del software sul data-set esempio.sas7bdat memorizzato nella cartella di installazione - sono riportate le stampe di tali file.

Ciascun file contiene una stampa, ad eccezione dei file stampa5.txt e stampa7.txt, che contengono entrambi due tabelle: stampa5.txt memoriz-

za il contenuto della stampe 5a e 5b mostrate nelle figure 5.21 e 5.22; stampa7.txt memorizza il contenuto della stampe 7a e 7b mostrate nelle figure 5.24 e 5.25 e stampa7.txt (per approfondimenti *cfr. paragrafi 6.5, 6.6, capitolo 6*).

Per migliorare la leggibilità delle stampe è conveniente:

- aprire tali file con Microsoft Word
- selezionare tutto il testo e convertirlo in SAS Monospace, punti 8.

Ciò renderà disponibili alcune informazioni, soprattutto per ciò che riguarda i file stampa5.txt stampa7.txt che, contenendo alcune tabelle scritte in formato SAS, risultano poco leggibili.

I **file excel** contengono le informazioni relative ai modelli per la presentazione sintetica degli errori campionari in formato excel e sono i seguenti:

- stampa5a.xls,
- stampa5b.xls,
- stampa7a.xls
- stampa7b.xls

Tali file vengono creati tramite le stampe numero 5 “Stampa modello 1” e numero 7 “Stampa modello 2”. In particolare:

- stampa 5a contiene le informazioni memorizzate nel data-set MODEL
- stampa5b.xls contiene una selezione delle variabili del data-set INTERP (*domst, perc, stima, errintp*);
- stampa7a.xls contiene le informazioni memorizzate nel data-set MODEL2;
- stampa7b.xls contiene una selezione delle variabili del data-set INTERP (*domst, perc, stima, errintp*).

Per approfondimento su tali *data-set* cfr. *capitolo 2, Sezione II*.

La tabella 5.1 mostra un esempio di file excel di output.

Tabella 5.1: il contenuto del file excel stampa5a.xls

DOMST	R2	A	B
PROV1	81.66504	7.221702	-1.59775
PROV2	83.82741	11.03019	-1.64744
PROV3	71.03369	11.35765	-1.70501
PROV4	84.88228	10.92052	-1.68725
PROV5	89.375	11.92888	-1.76303
TOTALE	88.30076	12.74548	-1.72878

Il **file di log** contiene le informazioni che appaiono nella finestra di log del SAS ed è il seguente:

- `genesees.log`

Il SAS durante le elaborazioni, permette la visualizzazione delle informazioni di esecuzione sulla finestra di *Log*. L'esecuzione del software Genesees crea un *Log*, che - data la sua lunghezza e complessità - viene registrato su un file esterno, nella cartella di output, con il nome "**genesees.log**".

Ciò è particolarmente utile nel caso di un messaggio di errore: le informazioni memorizzate sono visualizzabili anche successivamente. Per leggere il file `genesees.log` è necessario terminare l'esecuzione della procedura e uscire dal software.

I data-set di output sono i seguenti²:

a) Data-set di lavoro:

SAVEPAR creato per memorizzare parametri di input,

ERRORI_INPUT creato per memorizzare gli errori rilevati sull'input.

² La cartella di output scelta dall'utente corrisponde alla libreria "errori". Se, ad esempio, l'utente sceglie la cartella `c:\utente` - prendendo in considerazione il data-set di output STRATO - la procedura crea il data-set Sas di output "errori.strato" che corrisponde al file `c:\utente\STRATO.sas7bdat` (data-set sas v.8) registrato nella cartella `c:\utente`. Per semplificare l'esposizione successiva si farà riferimento ai data-set solo con il nome, senza l'estensione del file o la libreria di riferimento.

b) Data-set contenenti le informazioni relative a stime ed errori campionari:

STRATO, TOTALE, TOT_DIS0, TOT_DIS1

c) Data-set contenenti le informazioni relative a stime ed errori campionari utili ad elaborazioni successive:

WSTRATO, WTOTALE, WTOT_DIS0, WTOT_DIS1

d) Data-set contenenti informazioni sulla stratificazione e sul campione:

TAB1, UNIC

Per approfondire le informazioni contenute nei data-set sopra elencati, si può leggere il *capitolo 2 della Sezione II*.

SEZIONE II

**Approfondimenti sulla costruzione
dell'input e sui data-set di output della
funzione di calcolo delle Stime
e degli Errori di Genesees V. 3.0**

1. La costruzione del data-set di input

***Sintesi:** Nel paragrafo 1.1 sono presentate le variabili che devono essere contenute nel data-set di input e i parametri richiesti dalla funzione di calcolo delle stime e degli errori campionari. I vincoli che tali variabili devono rispettare sono successivamente introdotti nel paragrafo 1.2. Nei paragrafi 1.3 e 1.4 è mostrato come costruire tali variabili di input a seconda del tipo di stimatore adottato (1.3.1), in relazione al disegno di campionamento (1.3.2), in relazione a stime per domini pianificati (1.4.1) e non pianificati (1.4.2).*

1.1 Le variabili ed i parametri di input

Il *data-set* di input deve contenere diverse variabili, che possono essere raggruppate nelle seguenti tipologie:

- A. Variabili di interesse**
- B. Variabili di disegno**
- C. Variabili relative allo stimatore**
- D. Variabili relative al dominio di stima**

Inoltre l'utente deve scegliere i seguenti **parametri** di input:

- il tipo delle variabili di interesse: **qualitativo/quantitativo**;
- il **numero minimo di strati da aggregare** in un eventuale processo di *collassamento*;
- il **peso campionario**: *a livello di unità elementare o di cluster*.

Nel *paragrafo 1.1.1* vengono presentate le **variabili** del *data-set* di input, da

definire per il calcolo delle stime e degli errori campionari: il software richiede la presenza obbligatoria di alcune variabili di input e richiede specifici formati (in altre parole le variabili devono essere definite rigorosamente di tipo alfanumerico o di tipo numerico, come è di seguito indicato).

Nel *paragrafo 1.1.2* vengono infine descritti i **parametri** di input del software, specificandone il significato.

1.1.1 Le variabili di input

Attenzione! Il **nome** di tutte le variabili di input non può eccedere gli 8 caratteri!

A. Variabili di interesse

- 1) **Le variabili di interesse** sono le variabili oggetto di indagine, sulla base delle quali si costruiscono le stime dei parametri voluti. Attualmente la procedura è sviluppata per calcolare gli errori di campionamento delle stime dei totali riferiti alle variabili oggetto di indagine.

Caratteristiche delle variabili da costruire nel <i>data-set</i> di input:
Tipo: numerico
Valori da assumere: nessuna indicazione
Numero di variabili: 1 o più
Obbligatoria: almeno 1
Il nome delle variabili deve essere lungo al massimo 8 caratteri.

B. Variabili di disegno

- 2) **Tipo di disegno adottato:** identifica il tipo di disegno adottato e può assumere i valori alfanumerici “0” e “1”. Assume “1” solo per i disegni campionari ad uno stadio senza reimmissione delle unità e con probabilità di inclusione costante, “0” in tutti gli altri casi. Nel *data-set* di input, dunque, la variabile “Tipo di disegno” deve essere posta pari ad “1” per specificare che l’unità di campionamento è stata estratta secondo un disegno ad un unico stadio, senza reimmissione delle unità e con probabilità di inclusione costante; “Tipo di disegno” pari a “0” specifica disegni campionari diversi, sia ad

uno sia a due stadi, dove l'unità è stata selezionata con altri metodi di estrazione.

E' necessario aggiungere che il software permette di considerare disegni campionari *composti*. Un disegno *composto* definisce un campione le cui unità sono state estratte con disegni di campionamento differenti. Tramite la variabile "Tipo di disegno" è possibile indicare al software una prima distinzione relativa a due iniziali classi di disegni campionari, attribuendo alla variabile valori pari a "0" e pari ad "1". Nei successivi punti 3), 4) e 5) sono descritte le altre variabili attraverso le quali l'utente definisce completamente il disegno campionario.

L'uso di tale variabile è trattato dettagliatamente nel *paragrafo 1.3*.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input:
Tipo: alfanumerico
Valori da assumere: codice 0 o 1
Numero di variabili: 1
Obbligatoria.
Lunghezza: 1 carattere

- 3) **Unità primaria:** in un disegno a due o più stadi rappresenta il codice identificativo dell'unità primaria di campionamento. Le unità elementari nel *data-set* presentano il codice della unità primaria cui appartengono. Nei disegni ad uno stadio, poiché il software richiede obbligatoriamente di specificare tale variabile, l'utente può creare una variabile identica a quella definita nel punto 4) successivo.

Per ulteriori approfondimenti si veda il *paragrafo 1.3*.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico
Valori da assumere: qualsiasi
Numero di variabili: 1
Obbligatoria.

- 4) **Unità finale:** in un disegno a due o più stadi rappresenta il codice identificativo delle unità finali di campionamento. Le unità elementari appartenenti alla medesima unità finale di campionamento devono presentare, pertanto, lo stesso codice identificativo.

In un disegno ad un unico stadio, rappresenta l'unità primaria corrispon-

dente a quella finale di campionamento e dunque, a prescindere dal tipo di estrazione e dal valore assunto dalla variabile “Tipo di disegno”, per tutte le unità campionarie selezionate secondo un disegno ad un unico stadio i valori dei codici corrispondenti alle variabili “Unità primaria” e “Unità finale” saranno posti uguali.

Per approfondimento si veda il *paragrafo 1.3*.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico Valori da assumere: qualsiasi Numero di variabili: 1 Obbligatoria. Lunghezza: il numero può essere composto al massimo da 15 caratteri (lunghezza consigliata per compatibilità con la funzione di Riponderazione)

- 5) **Strato:** è il codice dello strato. Nel caso di indagini a più stadi di selezione la variabile strato si riferisce sempre alla stratificazione delle unità primarie. Le unità elementari appartenenti allo stesso strato devono presentare lo stesso codice identificativo dello strato. Per approfondimenti relativi alla definizione di questa variabile si veda l'introduzione al *paragrafo 1.3* (e i successivi rimandi).

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: alfanumerico Valori da assumere: qualsiasi Numero di variabili: 1 Obbligatoria.

- 6) **Il peso diretto:** la variabile indica il coefficiente diretto di riporto all'universo relativo all'unità elementare di campionamento. Nel caso di mancate risposte totali, il peso diretto deve essere stato precedentemente corretto per tenerne conto.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico Valori da assumere: qualsiasi Numero di variabili: 1 Obbligatoria.

C. Variabili relative allo stimatore

- 7) **Peso distanza:** è un peso da attribuire alla unità elementare di campionamento ed è utile per definire lo specifico stimatore adottato; per approfondimenti si veda il *paragrafo 1.3*.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input
Tipo: numerico Valori da assumere: qualsiasi Numero di variabili: 1 Obbligatoria.

- 8) **Variabili ausiliarie:** queste variabili corrispondono alle variabili ausiliarie utilizzate nello stimatore di calibrazione. Per approfondimenti sulla definizione di tali variabili nel *data-set* di input si veda l'introduzione del *paragrafo 1.3*. (e i successivi rimandi) e per gli aspetti metodologici correlati si veda l'*appendice A1*. E' opportuno ricordare che il software deriva tutti i principali e più utilizzati stimatori dalla teoria degli stimatori di calibrazione.

Caratteristiche delle variabili del <i>data-set</i> di input
Tipo: numerico Valori da assumere: qualsiasi Numero di variabili di interesse: 1 o più NON Obbligatorie

- 9) **Popolazioni pianificate utilizzate per lo stimatore:** la variabile serve ad individuare una partizione delle unità elementari; i relativi sottoinsiemi possono definirsi come *sottopopolazioni pianificate* (cfr. *paragrafo 1.3*) in quanto risultano sempre formati da strati o aggregazioni di strati.

Caratteristiche della variabile del <i>data-set</i> di input
Tipo: alfanumerico Valori da assumere: qualsiasi Numero di variabili: 1 Obbligatoria. Lunghezza: al massimo 15 caratteri (lunghezza consigliata per compatibilità con la funzione di Riponderazione)

- 10) **Peso finale:** la variabile indica il coefficiente finale di riporto all'universo. Può coincidere con il coefficiente iniziale, ad esempio nel caso in cui si utilizza uno stimatore di Horvitz-Thompson.

Caratteristiche della variabile del <i>data-set</i> di input
Tipo: numerico Valori da assumere: nessuna indicazione Numero di variabili: 1 Obbligatoria.

D. Variabili relative al dominio di stima

- 11) **Variabili di sottoclassi:** queste variabili servono a definire i domini di stima non pianificati, nel senso che servono a definire partizioni della popolazione rispetto alle quali interessano le stime finali (cfr. *paragrafo 1.3*). I domini di stima non pianificati sono sottoinsiemi della popolazione caratterizzati dal fatto che non tutte le unità di uno stesso strato appartengono allo stesso sottoinsieme della partizione. Ciascuna unità elementare presenta il codice identificativo della sottoclasse cui appartiene.

Caratteristiche delle variabili da costruire nel <i>data-set</i> di input:
Tipo: alfanumerico Valori da assumere: nessuna indicazione Numero di variabili: 1 o più NON Obbligatorie Lunghezza: al massimo 15 caratteri Il nome delle variabili deve essere lungo al massimo 8 caratteri.

- 12) **Dominio pianificato:** è il codice identificativo del dominio di stima pianificato: le modalità della variabile rappresentano i gruppi per i quali si desidera ottenere i totali delle variabili d'interesse a prescindere dall'uso di sottoclassi (cfr. *paragrafo 1.3*). Il dominio pianificato è caratterizzato dal rispetto della condizione che tutte le unità dello stesso strato appartengono ad uno ed un solo dominio pianificato – ovvero che un dominio pianificato corrisponde ad uno strato o è ottenibile anche come aggregazione di strati. Ciascuna unità elementare presenta il codice identificativo del dominio di stima cui appartiene.

Caratteristiche della variabile da costruire nel <i>data-set</i> di input:
Tipo: alfanumerico Valori da assumere: nessuna indicazione Numero di variabili: 1 Obbligatoria. Lunghezza: al massimo 15 caratteri

1.1.2. I parametri di input

☐ **Primo parametro: il tipo di variabili di interesse: qualitativo /quantitativo;**

Nella versione attuale, il *data-set* di input può contenere una o più variabili di interesse, tutte di tipo qualitativo o tutte di tipo quantitativo.

Per utilizzare il software e calcolare le stime e gli errori campionari, è necessario indicare il **tipo di variabile**, ovvero se le variabili sono di tipo qualitativo o quantitativo (tale parametro viene definito tramite l'uso dell'interfaccia – utilizzando il rispettivo bottone - quando si selezionano le “*Variabili di interesse*” - cfr. figura 5.5, *Sezione I*).

Nel caso in cui le variabili siano di tipo qualitativo, come ad esempio il sesso e la classe di età, il valore della variabile per ogni unità osservata corrisponde al valore della modalità della variabile assunta dalla medesima unità; tali modalità devono essere indicate con un valore numerico. In tal caso il software calcola l'errore di campionamento relativo alla stima di frequenza assoluta per ciascuna delle modalità della variabile stessa. Nel caso invece in cui le variabili siano di tipo quantitativo il software calcola l'errore di campionamento del valore quantitativo della stima del totale della variabile di interesse.

☐ **Secondo parametro: il numero minimo di strati per il collassamento**

L'utente può scegliere il **numero minimo di strati da aggregare in un eventuale processo di collassamento** (tale parametro viene definito tramite l'uso dell'interfaccia – utilizzando un campo editabile - quando si selezionano le “*Variabili di disegno*” - cfr. figura 5.8, *Sezione I*).

Per processo di *collassamento* si intende quel processo di aggregazione di strati con una unica unità primaria. Per tali strati non è infatti possibile calcolare la varianza: la stima della varianza riferita a quello strato corrisponderebbe ad un valore omesso, mentre la stima della varianza finale riferita ad un'eventuale partizione che include tale strato risulterebbe sottostimata.

Nel calcolare le stime e gli errori di campionamento il software - in auto-

matico - verifica se esistono strati con un'unica unità primaria e – sempre in automatico - prevede che per tali strati avvenga il collassamento, formando alcuni “superstrati”.

Per far ciò, gli strati che presentano una unica unità primaria sono aggregati tra loro (lasciando così inalterati gli strati su cui è possibile calcolare la varianza).

Se uno strato è aggregato ad un altro, risultano soddisfatte alcune condizioni, che il software verifica:

- gli strati aggregati devono essere formati da unità che appartengono alla stessa popolazione pianificata utilizzata per lo stimatore;
- per rispettare il livello di stima finale desiderato, gli strati aggregati devono essere formati da unità che appartengono anche allo stesso dominio di stima pianificato;
- per evitare di aggregare unità estratte secondo disegni diversi, gli strati aggregati devono essere formati da unità che presentano lo stesso valore della variabile “Tipo di disegno”.

Il software, nel formare i superstrati, aggrega per default due strati (è possibile variare tale valore di default tramite l'interfaccia: cfr figura 5.8, *capitolo 5, Sezione I*) e vengono aggregati eccezionalmente tre strati solo se, formando i superstrati, rimane un singolo strato aggregabile (ovvero che rispetta le condizioni di cui sopra rispetto a strati già aggregati).

❑ Terzo parametro: il tipo di peso

L'utente deve scegliere se il peso campionario utilizzato è *a livello di unità elementare* o *di cluster* (tale parametro viene definito tramite l'uso dell'interfaccia – utilizzando un bottone - quando si selezionano le “*Variabili relative allo stimatore*” - cfr. figura 5.10, *Sezione I*).

1.2 I vincoli sulle variabili

La costruzione del data-set di input richiede che siano rispettati i vincoli di coerenza e integrità tra le variabili.

Alcuni di questi vincoli sono controllati automaticamente dal software; altri, non essendo possibile un controllo automatico, devono essere a cura dell'utente. Nel caso in cui l'utente non abbia rispettato anche uno solo dei vincoli che il software controlla automaticamente, viene inviato un messaggio di avviso e **l'elaborazione è automaticamente interrotta**.

Di seguito vengono riportati i controlli effettuati automaticamente dal software. Prima di bloccare l'elaborazione il software scrive nella cartella di output un *data-set*, in cui registra l'incoerenza riscontrata.

Per maggiori informazioni si legga il *capitolo 2* in cui si analizza in dettaglio il *data-set* di output **errori_input.sas7bdat**.

I vincoli controllati dal software riguardano i punti:

- **Valori mancanti assunti dalle variabili di input**

Non possono esistere valori mancanti in alcuna delle variabili del *data-set* di input.

- **Unità primarie, unità finali ed unità elementari**

Ciascuna unità elementare è identificata da un codice assunto dalla variabile “unità primaria” e da un codice assunto dalla variabile “unità finale” di campionamento; il *data-set* di input non prevede codici identificativi per le unità elementari (il singolo record del *data-set*) e dunque non sempre sono identificabili chiavi univoche all'interno del *data-set*.

La procedura effettua il seguente controllo: il software controlla che le unità elementari con lo stesso codice di unità finale abbiano anche lo stesso codice di unità primaria. Il successivo schema è esemplificativo dei valori di un *data-set* di input in cui si leggono valori errati:

Schema 1:

Codice unità primaria	Codice unità finale errato
1	1
1	2
2	1
.....

- **Unità primarie e strati di appartenenza delle unità elementari**

A ciascuno strato appartengono più unità primarie di campionamento.

La procedura effettua il seguente controllo: il software controlla che le unità elementari con lo stesso codice di unità primaria abbiano anche lo stesso codice di strato. Il successivo schema è esemplificativo dei valori di un data-set di input in cui si leggono valori errati:

Schema 2:

Codice strato	Codice unità primaria errato
Str1	1
Str1	2
Str 2	1
.....

- **Popolazione pianificata utilizzata per lo stimatore, dominio di stima pianificato e strati di appartenenza delle unità elementari**

Sia la partizione specificata dalla variabile che indica la “Popolazione pianificata utilizzata per lo stimatore” sia la partizione che indica il “Dominio pianificato” definiscono sottoinsiemi, non necessariamente in ordine gerarchico tra loro, corrispondenti entrambi ad aggregazioni di strati.

Il software effettua il controllo tra ciascuna delle due variabili di cui sopra e la variabile strato: il software controlla che le unità elementari con lo stesso codice di strato abbiano anche lo stesso codice della variabile “Popolazione pianificata utilizza per lo stimatore”. Controlla analogamente che le unità elementari con lo stesso codice di strato abbiano anche lo stesso codice della variabile “Dominio pianificato”. Il successivo schema è esemplificativo dei valori di un data-set di input in cui si leggono valori errati (è mostrato un caso in cui il codice “str3” risulta errato in corrispondenza della variabile “Popolazione pianificata utilizzata per lo stimatore” ma non della variabile “Dominio pianificato”):

Schema 3:

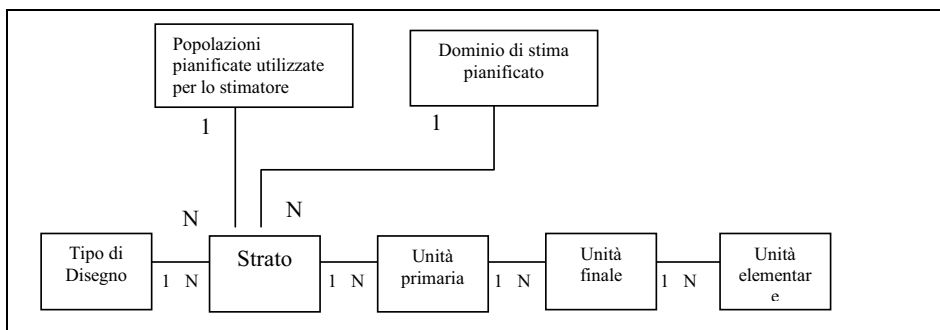
Codice pop.pian. Stimatore	Codice strato errato	Codice dominio pianificato
P1	Str1	D1
P1	Str1	D1
P1	Str2	D2
P2	Str3	D2
P3	Str3	D2
.....

- **Tipo di disegno adottato**

Il valore della variabile “tipo di disegno adottato” può essere solo “0” o “1” (valore alfanumerico). Il software controlla che le unità elementari con lo stesso codice della variabile “Strato”, abbiano il codice “Tipo di disegno” tutte pari ad “1” o tutte pari a “0”.

Per una migliore comprensione, si consulti lo schema 4, in cui vengono mostrate le relazioni usando concetti propri del modello Entità-Relazione (Chen P.P.S., 1976).

Schema 4: Le relazioni tra le variabili del data-set di input



A riguardo della relazione tra gli strati e le unità primarie, la relazione 1 a 1 indica che nel *data-set* possono esistere strati con un’unica unità primaria; il software per questi strati prevede in automatico un processo di *collassamento*, ampiamente descritto nel paragrafo precedente.

E' infine da prestare particolare attenzione al fatto che i suddetti vincoli si riferiscono a tutto il *data-set* SAS di input. Nel caso di indagini le cui unità sono estratte con disegni campionari diversi, tali vincoli sono dunque validi per i codici di tutte le unità elementari del *data-set*.

1.3 Definizione delle variabili di input in relazione alla strategia campionaria

Il software implementa diversi stimatori della varianza campionaria, ciascuno dei quali produce una stima corretta o approssimativamente corretta per un particolare stimatore del parametro di interesse e del disegno di campionamento adottato.

Al fine di ottenere la stima della varianza campionaria l'utente deve definire opportunamente alcune variabili del *data-set* di input (in seguito indicato con il nome INP).

La tabella 1.1 indica le variabili alle quali l'utente dovrà porre attenzione. A tali variabili per semplicità espositiva sono stati assegnati dei nomi di sintesi.

Tabella 1.1: variabili del data-set di input INP usate per definire la strategia campionaria adottata

Variabili di input (paragrafo 1.1.1)	Nome sintetico della variabile
Tipo di disegno adottato	TIPO_DIS
Unità primaria	UNITA_1
Unità finale	UNITA_2
Strato	STRATO
Peso diretto (corretto per mancata risposta totale)	COEF_DIR
Peso finale	COEF_FIN
Variabili ausiliarie	X1, ..., Xj, ..., XJ
Popolazione pianificata utilizzata per lo stimatore	POP_PIAN
Peso distanza	CK

I successivi *paragrafi 1.3.1* e *1.3.2* approfondiscono gli aspetti legati alla definizione delle variabili necessarie ad identificare lo stimatore della varianza per un dato stimatore del parametro; nel *paragrafo 1.3.3* si descrivono le caratteristiche delle variabili di input necessarie per specificare il disegno campionario che ha dato origine ai coefficienti finali presenti nel *data-set* INP.

La tabella 1.2 indica quali variabili di input sono legate alla definizione dello stimatore campionario e quali variabili sono connesse al disegno.

Tabella 1.2 – Variabili di input che utilizzate per definire la strategia campionaria

	TIPO_DIS	CK	UNITA_2	UNITA_1	POP_PIAN	STRATO	Xj
Stimatore		X			X	X	X
Disegno	X		X	X		X	

E' necessario sottolineare che le variabili POP_PIAN, $X_1, \dots, X_j, \dots, X_J$ e CK sono già definite correttamente se il *data-set* INP risulta essere l'output della funzione di *Riponderazione*.

1.3.1 Definizione delle variabili di input per un dato stimatore

La funzione *Stime ed Errori* del software adotta gli stimatori corretti o approssimativamente corretti della varianza campionaria più noti in letteratura (cfr. *appendice A.3*) per la classe degli stimatori di *ponderazione vincolata* o *calibrazione* del parametro totale. A tale famiglia appartengono tutti i principali stimatori che utilizzano informazioni ausiliarie, quali gli stimatori del *rapporto*, *rapporto post-stratificato*, *raking* e *regressione generalizzata* (per gli aspetti metodologici cfr. *appendice A.1*). Tale classe può essere estesa includendo anche lo stimatore di *Horwitz-Thompson* e *espansione*.

Per indicare al software il processo di stima che ha generato i coefficienti finali di input (COEF_FIN), l'utente deve agire sulle variabili del *data-set* INP, denominate POP_PIAN, $X_1, \dots, X_j, \dots, X_J$, CK e STRATO (cfr. tabella 1.1).

I successivi paragrafi affrontano i seguenti aspetti:

- i *paragrafi 1.3.1.1 e 1.3.1.2* descrivono le caratteristiche delle variabili POP_PIAN e $X_1, \dots, X_j, \dots, X_J$ per individuare il *gruppo di riferimento del modello* (per approfondimenti cfr. *appendice A.1.1*). Tali paragrafi sono indirizzati ad utenti che devono ottenere le stime della varianza per stimatori complessi quali *raking generalizzato*, *regressione generalizzata* e *ponderazione vincolata*, che utilizzano più di un totale noto di riferimento;
- il *paragrafo 1.3.1.3* pone l'attenzione sulla variabile CK in relazione al

livello del modello dello stimatore che ha determinato i coefficienti finali di riporto (per approfondimenti cfr. *appendice A.1.2*). Tale paragrafo è indirizzato in particolare agli utenti che devono stimare la varianza per stimatori definiti a *livello di cluster*. In generale non è necessaria la lettura del paragrafo per gli utenti che devono stimare la varianza degli stimatori di *Horvitz-Thompson*, del *rapporto* e *ratio raking*.

- Il *paragrafo 1.3.1.4* suggerisce i valori che devono assumere le variabili POP_PIAN, $X_1, \dots, X_j, \dots, X_J$, CK e STRATO affinché si specifichi il *tipo di modello* (per approfondimenti cfr. *appendice A.1.3*). La lettura di questo solo paragrafo è sufficiente per gli utenti che devono ottenere le stime della varianza per gli stimatori di *Horvitz-Thompson*, del *rapporto* e *ratio raking*.

Per gli utenti che fanno uso della funzione *Stime ed Errori* dopo aver utilizzato la funzione *Riponderazione*, le variabili POP_PIAN, $X_1, \dots, X_j, \dots, X_J$ e CK sono già definite in modo corretto. E' sufficiente, pertanto, considerare la sola variabile STRATO e leggere il *paragrafo 1.3.3*.

1.3.1.1 GRUPPO DI RIFERIMENTO DEL MODELLO

I *gruppi di riferimento del modello* (*model groups*)³, in base ai quali sono definiti gli stimatori regressione generalizzata e di ponderazione vincolata o calibrazione, sono delle particolari sottopopolazioni della popolazione di interesse per le quali si conoscono i totali (detti *totali noti*) di alcune variabili ausiliarie utilizzate per la costruzione dello stimatore stesso (per approfondimenti cfr. *appendice A.1.1*).

Per indicare al software quali sono i gruppi di riferimento che sono stati considerati dallo stimatore per il quale si vuole calcolare la varianza campionaria, è necessario strutturare il *data-set* INP nel modo opportuno. A tal fine è tuttavia necessario formulare una premessa.

³ L'espressione "gruppo di riferimento del modello" ha origine dalla terminologia adottata per lo stimatore di regressione generalizzata, in cui il gruppo di riferimento è un sottogruppo del campione per il quale si stima il modello di regressione.

Data la popolazione dalla quale si estrae il campione, si può definire una sottopopolazione in relazione alla stratificazione del disegno di campionamento adottato. In tal caso si possono individuare due tipi di sottopopolazioni: le *sottopopolazioni pianificate* e le *sottopopolazioni non pianificate*.

Le sottopopolazioni pianificate, definibili quando si adotta un disegno stratificato, sono costruite in modo tale da coincidere con uno o più strati del disegno. In tal caso, dato uno strato o un insieme di strati, tutte le unità dello strato o dell'insieme di strati appartengono ad una ed una sola sottopopolazione pianificata.

Le sottopopolazioni non pianificate, invece, sono costruite in modo tale che le unità di un generico strato appartengono solo in parte ad una generica sottopopolazione. In generale, se la sottopopolazione non pianificata è costituita da unità provenienti da strati diversi, è importante che, per almeno uno strato, non siano presenti tutte le unità nella sottopopolazione perché questa possa definirsi non pianificata.

I gruppi di riferimento, essendo sottopopolazioni in cui sono noti dei totali per alcune variabili ausiliarie, si possono classificare come pianificati e non pianificati. In particolare, procedendo ad una descrizione più dettagliata si hanno le tre categorie seguenti:

- (i) sottopopolazioni pianificate;
- (ii) sottopopolazioni non pianificate definite all'interno di strati o aggregazioni di strati;
- (iii) sottopopolazioni non pianificate.

Per quanto riguarda il caso (i) si ha che ciascun gruppo di riferimento può essere formato da:

- (A1) tutte le unità appartenenti ad un singolo strato del disegno;
- (A2) tutte le unità appartenenti ad un'aggregazione di strati del disegno;
- (A3) l'intera popolazione di riferimento (in questo caso si ha un'unica sottopopolazione pianificata che coincide con la popolazione stessa).

Considerando invece il caso (ii), ciascun gruppo di riferimento può contenere:

- (B1) una parte delle unità contenute in uno strato;
- (B2) una parte delle unità contenute in una aggregazione di strati;

Infine per il caso (iii), ciascun gruppo di riferimento deve essere composto da:

- (B3) una parte delle unità contenute nella popolazione (che non coincide con l'insieme completo di unità contenute in uno strato o in una aggregazione di strati).

I D gruppi di riferimento, che costituiscono una partizione della popolazione, formata secondo uno dei sei criteri sopra illustrati, risultano allora pari a:

- (A1) $D=H$, in cui H rappresenta il numero complessivo degli strati;
- (A2) $D=H_G (<H)$, in H_G è il numero degli insiemi di strati aggregati;
- (A3) $D=1$;
- (B1) $D=H \times Q$, in cui Q è il numero delle sottopopolazioni non pianificate presenti all'interno dello strato. Queste sottopopolazioni devono essere definite allo stesso modo in ciascuno strato;
- (B2) $D=H_G \times Q$, in cui Q è il numero delle sottopopolazioni non pianificate presenti all'interno di una aggregazione di strati. Queste sottopopolazioni devono essere definite allo stesso modo in ciascuna aggregazione di strati;
- (B3) $D=Q$, in cui Q è il numero delle sottopopolazioni non pianificate presenti all'interno della popolazione di studio.

Definendo i gruppi di riferimento secondo uno dei sei punti precedenti, si può osservare che le partizioni ottenute con i punti (A1), (A2) e (A3) rappresentano casi particolari rispettivamente dei punti (B1), (B2) e (B3) quando si ha $Q=1$.

Tenendo in considerazione queste tipologie di gruppi di riferimento, il software offre la possibilità di individuare i gruppi di riferimento del modello definendo in modo opportuno le variabili POP_PIAN e $X_1, \dots, X_j, \dots, X_J$. Nei successivi sottoparagrafi sono descritte le varie possibilità in funzione del tipo di totali noti a disposizione.

□ **Stimatore definito su una sola variabile ausiliaria e una sola-partizione dell'universo in gruppi di riferimento**

Si consideri il caso in cui si vuole stimare la varianza di uno stimatore di ponderazione vincolata, che tiene conto dei totali noti di una variabile

ausiliaria X (per approfondimenti sul tipo di variabile ausiliaria cfr. *appendice A.4.2*) su un insieme di D gruppi di riferimento definiti combinando una variabile di stratificazione s (con modalità $h=1, \dots, H$)⁴ e una variabile v (con modalità $q=1, \dots, Q$), che non contribuisce alla stratificazione del disegno. I $D (=H \times Q)$ ⁵ gruppi costituiscono una partizione P della popolazione. Gli stimatori che presentano una tale struttura di totali noti sono quello del *rapporto separato* (in cui $Q=1$), del *rapporto post-stratificato* (in cui $H=1$), del *rapporto post-stratificato separato* e dello stimatore del *rapporto post-stratificato combinato* (in cui $H=1$). Rientrano in questa classe, considerando il caso $D=1$, anche lo stimatore di *Hàjek* (stimatore del rapporto che utilizza la numerosità della popolazione come totale noto), del *rapporto semplice*, *rapporto combinato* e di *regressione semplice*.

Per indicare al software quali sono i gruppi di riferimento è necessario definire correttamente le variabili POP_PIAN e $X_1, \dots, X_j, \dots, X_J$ nel *data-set* INP (cfr. tabella 1.1).

A tale riguardo si consideri l'esempio seguente:

Esempio 1.1:

Si consideri una strategia di campionamento in cui sia stato estratto un campione di individui, da una popolazione stratificata secondo la variabile sesso (variabile s). I coefficienti di riporto finali (COEF_FIN) dello stimatore impiegato riportano al totale noto degli individui per ognuna delle combinazioni delle modalità delle variabili sesso e classe di età (variabile v).

Nella tabella 1.3. sono descritte le due variabili. I gruppi di riferimento sono in totale 8.

Tabella 1.3 – Descrizione delle variabili che definiscono lo stimatore nell'esempio 1.1

Variabile	Modalità della variabile	Numero modalità	Simbolo	Numero delle modalità (simbolo)
Sesso	Uomo; Donna.	2	S	S
Classe di età	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre.	4	V	Q

⁴

Per brevità non si descrive il caso in cui la variabile s è costituita da modalità che sono aggregazioni degli strati del disegno. Questo caso è facilmente ricavabile dalle considerazioni sviluppate nel paragrafo e viene ripreso nell'appendice A.4.

⁵

In generale i gruppi di riferimento si possono identificare anche combinando una variabile che rappresenta una aggregazione degli strati del disegno (si veda appendice A.4).

Per considerare correttamente questa struttura di totali noti alla base di coefficienti finali di riporto il data-set INP può essere costruito secondo diverse alternative.

La prima alternativa, denominata schema A, è illustrata nella figura 1.1.

Figura 1.1 – Costruzione del data-set INP secondo lo schema A

VIEWTABLE: Work.Esempioinp											
	cod	A	X1	X2	X3	X4	X5	X6	X7	X8	...
1	1 cost		0	0	0	0	0	1	0	0	...
2	2 cost		0	0	0	1	0	0	0	0	...
3	3 cost		0	0	0	0	0	0	0	1	...
4	4 cost		1	0	0	0	0	0	0	0	...
5	5 ...		1	0	0	0	0	0	0	0	...

Nel data-set per ciascun record sono presenti, tra le altre, la variabile A e le variabili X1,...,X8. Per quanto riguarda la variabile A essa presenta su tutti i record un’unica modalità (definita arbitrariamente dall’utente). Per quanto riguarda le variabili X1, ..., X8 queste individuano gli otto gruppi di riferimento. La corrispondenza variabile Xj – gruppo di riferimento è definita dall’utente. Ad esempio, si possono identificare nelle prime quattro variabili le quattro classi di età degli uomini e nelle seconde quattro, le classi di età delle donne. In questo caso la variabile X1 individua il gruppo di riferimento descritto dalla combinazione delle modalità uomo e 0-14 anni.

Il criterio per assegnare i valori alle variabili Xj è il seguente: se il record appartiene al gruppo di riferimento individuato dalla generica variabile Xj, tale variabile assume il valore osservato della variabile x (in questo caso il valore è 1); se il record non appartiene al gruppo di riferimento individuato da Xj tale variabile assume valore nullo. Ad esempio, il primo record della figura 1.1 è una donna con età 15-34 anni.

La seconda alternativa, denominata schema B, prevede che gruppo di riferimento sia individuato dalla combinazione della modalità di riga della variabile POP_PLAN e della variabile ausiliaria presente in colonna. Nell’esempio riportato nella figura 1.2., la cella è identificata dalla combinazione della modalità di riga della variabile sesso, che assume il ruolo della variabile POP_PLAN, con una delle colonne identificate dalle variabili X1, ..., X4, che identificano le modalità della variabile classe di età. Ad esempio la variabile X1 può identificare la classe 0-14 e così via per le altre variabili.

Nella figura 1.2 il primo record rappresenta pertanto una donna con età 15-34 anni.

Figura 1.2 – Costruzione del data-set INP secondo lo schema B

VIEWTABLE: Work.Esempioinp							
	cod	sess	X1	X2	X3	X4	...
1	1	donna	0	1	0	0	...
2	2	uomo	0	0	0	1	...
3	3	donna	0	0	0	1	...
4	4	uomo	1	0	0	0	...
5	5	...	1	0	0	0	...

In generale se il record presenta la classe di età individuata dalla generica variabile X_j , tale variabile assume il valore osservato della variabile x (in questo caso il valore è 1); altrimenti la generica variabile X_j assume valore nullo.

Alcune importanti indicazioni che si possono trarre da questo esempio, e che sono sempre valide nella costruzione dei *data-set*, sono le seguenti:

- la variabile POP_PIAN può assumere come modalità solo quelle che definiscono gli strati (come nello schema B dell'esempio 1.1) o modalità di variabili che rappresentano aggregazioni di strati (in particolare lo schema A rappresenta un caso estremo di aggregazione di tutti gli strati del disegno). Ciascuna modalità di POP_PIAN identifica, pertanto, una sottopopolazione pianificata;
- le variabili X_1, \dots, X_J , individuano le modalità che, combinate con quelle della variabile POP_PIAN, definiscono i gruppi di riferimento dello stimatore di ponderazione vincolata. Tali variabili da una parte devono essere definite attraverso le modalità delle variabili che definiscono i gruppi di riferimento, ma non rientrano nella definizione della stratificazione; dall'altra possono essere definite considerando anche le modalità delle variabili che contribuiscono a definire la stratificazione del disegno (è questo il caso dello schema A nell'esempio 1.1). Ciascuna variabile X_j può identificare, pertanto, una qualsiasi sottopopolazione, pianificata o non pianificata.

Le regole generali per la costruzione del *data-set* INP (rispettate nell'esempio 1.1) secondo lo schema A sono descritte nell'elenco seguente e nella tabella 1.4.

- Caratteristiche del data-set INP secondo lo schema A;
 - la variabile POP_PIAN risulta costante per ciascun record del *data-set*;
 - il numero delle variabili X_j corrisponde ai D gruppi di riferimento. Ogni variabile X_j ($j=1, \dots, d, \dots, D$) identifica uno specifico gruppo di riferimento. Per ciascun record solo una di queste variabili X_j assume il valore della variabile x osservato sul record stesso, mentre le altre sono nulle. La variabile che presenta il valore di x è quella che identifica il gruppo di riferimento a cui appartiene il record stesso.

Tabella 1.4 - Descrizione dello Schema A: definizione del data-set INP con una variabile ausiliaria x e una partizione P con D gruppi di riferimento. Esempio per il record appartenente al j -esimo gruppo di riferimento.

POP_PIAN	X1	...	$X_j(j=d)$...	$X_J(J=D)$
...
...
Costante	0	...	x	...	0
...
...

Le regole per la costruzione del *data-set* INP secondo lo schema B sono presentate nell'elenco seguente e nella tabella 1.5.

- Caratteristiche del data-set INP secondo lo schema B;
 - la variabile POP_PIAN presenta per ciascun record la modalità della variabile che definisce una sottopopolazione pianificata (uno strato o un'aggregazione di strati), a cui appartiene il record stesso;
 - sono presenti una serie di variabili X_j , ciascuna delle quali coincide con una modalità della variabile che definisce i gruppi di riferimento ma non rientra nella definizione della stratificazione del disegno. Considerando una generica X_j , questa assume il valore della variabile x osservata sul record stesso se il record appartiene al gruppo di riferimento identificato dalla combinazione della modalità assunta dalla variabile POP_PIAN e da quella individuata dalla stessa variabile X_j , altrimenti la variabile X_j assume valore nullo.

Tabella 1.5 - Descrizione dello Schema B: definizione del data-set INP con una variabile ausiliaria x e una partizione P con D gruppi di riferimento. Esempio per il record i appartenente al d -esimo ($d=\{h;q\}$) gruppo di riferimento.

POP_PIAN	X1	...	$X_j(j=q)$...	$X_J(J=Q)$
...
...
H	0	...	x	...	0
...
...

In alcune occasioni si rende disponibile una terza alternativa, denominata schema C, per costruire i gruppi di riferimento. Ciò avviene quando i gruppi sono definiti con una variabile che individua le sottopopolazioni pianificate s come risultato della combinazione di due o più variabili s_1, \dots, s_R . Di seguito è presentato un esempio con $R=2$.

Esempio 1.2:

Si consideri una strategia di campionamento in cui sia stato estratto un campione di individui da una popolazione stratificata secondo la variabile combinata (variabile s) sesso (variabile s_1) e ripartizione geografica di residenza dell'unità campionaria (variabile s_2). I coefficienti di riporto finali (COEF_FIN) dello stimatore impiegato riportano al totale noto degli individui per ognuna delle combinazioni delle modalità delle variabili sesso, ripartizione geografica e classe di età (variabile v).

Nella tabella 1.6 sono descritte le tre variabili. I gruppi di riferimento sono in totale 24.

Tabella 1.6 – Descrizione delle variabili che definiscono lo stimatore nell'esempio 2

Variabile	Modalità della variabile	Numero modalità	Simbolo	Numero delle modalità (simbolo)
Sesso × Ripartizione geografica	Uomo×Nord; Uomo×Centro; Uomo×Sud; Donna×Nord; Donna×Centro; Donna×Sud;	6	s	S
Classe di età	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre.	4	v	Q

Per una tale struttura di totali noti, seguendo lo schema A devono definirsi 24 variabili X_j , mentre attraverso lo schema B si devono costruire 6 variabili X_j .

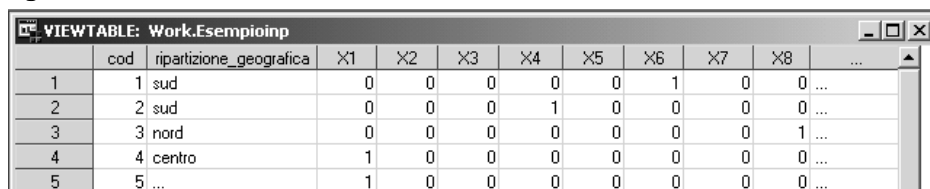
Tuttavia, avendo a disposizione una variabile di stratificazione combinata con due variabili che rientrano nella definizione dei gruppi di riferimento si può procedere ad una terza alternativa (schema C) per la costruzione dei data-set di input.

Si sceglie una tra le variabili sesso e ripartizione geografica. La variabile scelta, ad esempio la ripartizione geografica, assume il ruolo della variabile POP_PLAN. Si costruisce quindi una nuova variabile, combinando la variabile non scelta precedentemente, nell'esempio la variabile sesso, con la variabile classe di età.

Ad esempio le variabili X_1, \dots, X_4 individuano le quattro classi di età per gli uomini, mentre le variabili X_5, \dots, X_8 , sono relative alle quattro classi di età per le donne. La corrispondenza variabile X_j – combinazione delle modalità sesso e classe di età è definita dall'utente.

La figura 1.3., mostra come si presenta il data-set INP. Il primo record è relativo ad una donna con età 15-34 anni residente nel sud.

Figura 1.3 - Costruzione del data-set INP secondo lo schema C



	cod	ripartizione_geografica	X1	X2	X3	X4	X5	X6	X7	X8	...
1	1	sud	0	0	0	0	0	1	0	0	...
2	2	sud	0	0	0	1	0	0	0	0	...
3	3	nord	0	0	0	0	0	0	0	1	...
4	4	centro	1	0	0	0	0	0	0	0	...
5	5	...	1	0	0	0	0	0	0	0	...

In termini generali, quando la variabile che definisce le sottopopolazioni pianificate alla base dei gruppi di riferimento è composta da un insieme di due o più variabili, per definire i due data-set con lo schema C, è necessario suddividere preventivamente queste variabili in due classi. Attraverso la combinazione delle modalità delle variabili appartenenti alla prima classe si definiscono le S_1 modalità della variabile POP_PIAN. Attraverso la combinazione delle S_2 modalità della seconda classe di variabili, e le Q modalità di una eventuale variabile che non rientra nella definizione della stratificazione del disegno, si determinano le $Q \times S_2$ variabili X_j .

Come per lo schema B, la variabile che coincide con POP_PIAN deve quindi rappresentare una combinazione di modalità delle variabili che contribuiscono alla stratificazione oppure la combinazione di aggregazione di

tali modalità. In ogni caso viene rispettato il principio per cui ogni modalità della variabile POP_PIAN identifica una sottopopolazione pianificata.

Nella tabella 1.7 sono illustrate sinteticamente le regole per la costruzione del *data-set* INP secondo lo schema C.

Tabella 1.7 - Descrizione dello Schema C: definizione del data-set INP con una variabile ausiliaria x e una partizione P con D gruppi di riferimento. Stratificazione del disegno secondo le modalità di una variabile s ottenuta come combinazione delle variabili s_1 (con modalità $h_1=1, \dots, H_1$) e s_2 (con modalità $h_2=1, \dots, H_2$). Esempio per il record i appartenente al d -esimo ($d=(h_1;h_2;q)$) gruppo di riferimento.

POP_PIAN	X1	...	Xj(j=q)	...	XJ(J=Q)
...
...
H	0	...	x	...	0
...
...

□ Stimatore definito su più variabili ausiliarie e su più partizioni dell’universo in gruppi di riferimento

Il processo di calibrazione che ha originato i coefficienti finali di riporto può aver fatto ricorso ai totali noti di più d’una variabile ausiliaria e per più partizioni dell’universo in gruppi di riferimento. Per alcuni stimatori, ad esempio, si possono avere a disposizione i totali di una variabile ausiliaria per diverse partizioni, e partizioni sulle quali sono noti i totali di diverse variabili ausiliarie. Stimatori, noti in letteratura, che presentano questa struttura più generale dei totali noti sono il *ratio raking* (una variabile ausiliaria e due partizioni in gruppi di riferimento), il *raking generalizzato* e *regressione generalizzata* o, ancora più in generale, gli stimatori di *ponderazione vincolata*.

Assumendo, pertanto, di avere a disposizione $x_1, \dots, x_p, \dots, x_T$, variabili ausiliarie per le quali sono noti i totali su varie partizioni in gruppi di riferimento della popolazione obiettivo, la costruzione del *data-set* di input può seguire uno dei tre diversi schemi introdotti nel paragrafo precedente.

- Caratteristiche del data-set INP secondo lo schema A;
- la variabile POP_PIAN risulta costante per ciascun record del *data-set*;
 - per ogni variabile x_f e per la partizione associata in cui sono noti i totali si crea un insieme di variabili X_j . Il numero delle variabili X_j nell'insieme è dato dal numero di gruppi di riferimento che identificano la partizione. Per ciascun record solo una di queste variabili X_j dell'insieme assume il valore della variabile x_f osservato sul record stesso, mentre le altre sono nulle. La variabile che presenta il valore di x_f è quella che identifica il gruppo di riferimento a cui appartiene il record stesso;
 - si forma un insieme di variabili X_j per ogni partizione e ogni variabile x_f in cui sono noti i totali.

Una descrizione dello schema A si può ricavare dalla tabella 1.4. In questo caso si considera una sola coppia partizione – variabile ausiliaria, per la quale si conoscono i totali di popolazione.

Per quanto riguarda lo schema B, la sua applicazione è possibile solo quando:

- la variabile che definisce le sottopopolazioni pianificate rientra nella definizione dei gruppi di riferimento di tutte le variabili ausiliarie.

La costruzione dell' input segue in pratica le istruzioni illustrate nel paragrafo precedente (relative allo schema B), in cui si ha a disposizione una sola variabile ausiliaria.

La tabella 1.5 descrive le caratteristiche del *data-set* di input.

Relativamente all'applicazione dello schema C, questa è resa possibile quando:

- le modalità della variabile che definisce le sottopopolazioni pianificate sono ottenibili come combinazione di due o più variabili;
- è possibile individuare un sottoinsieme di variabili (tra quelle che definiscono le sottopopolazioni pianificate) che contribuiscono con le loro combinazioni di modalità a definire tutte le partizioni prese

in considerazione dal processo di calibrazione (tale sottoinsieme di variabili definisce a sua volta delle sottopopolazioni pianificate);

- le combinazioni delle modalità delle variabili *comuni* (o loro sottoinsiemi) sono utilizzate per definire la variabile POP_INP;
- le altre variabili che definiscono le sottopopolazioni pianificate alla base dei gruppi di riferimento, contribuiscono alla definizione delle variabili X_j nel *data-set* INP come avviene nello schema A e B.

La costruzione del *data-set* secondo lo schema C è illustrata sinteticamente nella tabella 1.7. In tale tabella si fa riferimento ad una coppia partizione – variabile ausiliaria, per la quale si conoscono i totali di popolazione.

Per ulteriori approfondimenti relativi alla costruzione dei *data-set* di input si rimanda all'*appendice A.4*.

1.3.1.2 SCELTA DELLO SCHEMA DI COSTRUZIONE DEL DATA-SET INP

Prima di presentare alcune indicazioni per la scelta di uno dei tre schemi per la costruzione del *data-set* di input, bisogna distinguere il caso in cui l'utente utilizza un *data-set* sul quale è stata applicata la funzione di *Riponderazione* del software oppure quando il *data-set* con i coefficienti finali di riporto ha una diversa origine. Nel primo caso, infatti, il *data-set* evidentemente è già stato costruito seguendo uno dei tre schemi, ed è quindi pronto per l'applicazione della funzione *Stime ed Errori*. Nel secondo caso se si deve operare una scelta tra le tre alternative si devono tenere in considerazione i punti seguenti:

- i vincoli operativi;
- i vantaggi e gli svantaggi connessi con l'efficienza computazionale del software;
- la possibilità di applicare il metodo del *collassamento* degli strati per la stima della varianza (cfr. *paragrafi 1.1.2 e 2.8*);

Questi punti sono descritti nella tabella 1.8.

Tabella 1.8 – Vincoli, vantaggi e svantaggi dei diversi schemi di definizione del data-set di input

Metodi di formazione del data set		
Schema A	Schema B	Schema C
Vincoli		
Non esistono vincoli sulla variabile che definisce le sottopopolazioni pianificate.	La variabile che definisce le sottopopolazioni pianificate deve essere comune a tutte le partizioni in gruppi di riferimento.	La variabile che definisce le sottopopolazioni pianificate deve essere composta da due o più variabili. Almeno una variabile o una classe di variabili che compone la variabile che definisce la sottopopolazione pianificata deve essere comune a tutte le partizioni in gruppi di riferimento.
Vantaggi		
La costruzione delle variabili di input che definiscono i gruppi di riferimento è diretta. Si può sempre applicare il metodo del collassamento degli strati.	Nella costruzione delle variabili di input che definiscono i gruppi di riferimento si richiede solo la suddivisione tra la variabile che definisce le sottopopolazioni pianificate dalle altre variabili. Per campioni di grandi dimensioni quando le modalità della variabile che definiscono le sottopopolazioni pianificate sono numerose può essere più efficiente dello schema A.	Per campioni di grandi dimensioni ed indagini multiobiettivo questa impostazione garantisce in genere una migliore efficienza computazionale quando la suddivisione delle variabili che definiscono POP_PIAN e TXj (o Xj) determina un equilibrio tra il numero delle modalità della variabile POP_PIAN ed il numero delle variabili TXj (o Xj).
Svantaggi		
Per campioni di grandi dimensioni, ed indagini multiobiettivo si possono presentare problemi di ordine computazionale causati dal numero elevato di variabili ausiliarie TXj e Xj.	Per campioni di grandi dimensioni, lo schema può risultare computazionalmente meno efficiente dello schema C a causa di un eventuale numero elevato di modalità della variabile POP_PIAN. Il metodo del collassamento degli strati a volte non si può applicare.	Il metodo può richiedere alcune operazioni preventive per suddividere la variabile che definisce le sottopopolazioni pianificate in due classi di variabili. Il metodo del collassamento degli strati a volte non si può applicare.

1.3.1.3 LIVELLO DEL MODELLO

Il concetto di *livello del modello* indica il tipo di unità utilizzata nella formulazione del modello di regressione sottostante allo stimatore utilizzato. In particolare, il software consente di formulare il modello sia a *livello di unità elementare* che a *livello di cluster di unità elementari*.

E' importante ricordare che nella versione attuale del software la funzione *Riponderazione* non consente la calibrazione dei coefficienti diretti di riporto con il modello a livello di cluster.

□ **Modello a livello di unità elementare**

Il modello *a livello di unità elementare* si può impostare con qualsiasi disegno campionario. Nella costruzione del *data-set* di input è necessario fare attenzione alle variabili che definiscono il disegno campionario (cfr. il *paragrafo 1.3.2*) e alle variabili COEF_FIN e CK. Queste ultime due devono essere definite nel modo seguente:

- ogni record può presentare un valore del coefficiente di riporto finale (COEF_FIN) diverso;
- non esistono particolari vincoli sulla variabile CK. I valori che può assumere tale variabile per ciascun record sono descritti nel *paragrafo 1.3.1.4*.

□ **Modello a livello di cluster**

Il modello *a livello di cluster* si può impostare con disegni ad uno o più stadi di selezione in cui le unità finali di campionamento sono costituite da grappoli (clusters) di unità elementari. Le variabili sulle quali bisogna porre l'attenzione sono le variabili che definiscono il disegno di campionamento (cfr. il *paragrafo 1.3.2*) e le variabili COEF_FIN e CK. Considerando queste ultime due variabili bisogna distinguere due casi:

- i record del *data-set* INP rappresentano unità elementari;
- i record del *data-set* INP rappresentano cluster di unità elementari;

Nel primo caso devono essere soddisfatte le seguenti condizioni:

- tutti i record di un medesimo cluster devono presentare lo stesso

valore della variabile COEF_FIN (coefficiente finale) e lo stesso valore della variabile UNITA_2;

- tutti i record di un medesimo cluster devono avere lo stesso valore della variabile CK.

Nel secondo caso:

- tutte le informazioni del record si devono riferire al cluster.

Costruire il *data-set* con i record che rappresentano cluster di unità elementari a volte può essere una scelta obbligata. Tale situazione si presenta quando si conoscono i valori delle variabili ausiliarie solo per i grappoli e non per le singole unità elementari che vi appartengono.

Infine, alcuni suggerimenti per la definizione della variabile CK per specificare il modello a livello di cluster sono forniti al termine del *paragrafo 1.3.1.4*.

1.3.1.4 TIPO DI MODELLO

La funzione *Stime ed Errori* permette di stimare la varianza corretta o approssimativamente corretta degli stimatori di ponderazione vincolata o calibrazione del parametro totale. Per indicare al software quale particolare stimatore ha dato origine ai coefficienti finali di input (COEF_FIN) è necessario che alcune variabili del *data-set* INP presentino determinati valori. Con tale operazione si definisce il *tipo di modello* utilizzato nel processo di calibrazione (per approfondimenti cfr. *appendice A.1.3*).

Di seguito sono illustrate le principali caratteristiche di INP per alcuni importanti stimatori ottenuti attraverso la calibrazione definita a livello di unità elementare. Per gli analoghi stimatori ottenuti con la calibrazione a livello di cluster si veda la parte *Stimatori a livello di cluster*, al termine del paragrafo.

Le espressioni relative agli stimatori trattati in questo paragrafo e le loro varianze sono riportate nell'*appendice A.1. e A.3*.

Per gli utenti che fanno uso della funzione *Stime ed Errori* dopo aver utilizzato la funzione *Riponderazione*, è sufficiente porre l'attenzione sulla sola variabile STRATO, che non è richiesta per il lancio di tale funzione.

Per gli utenti che devono stimare la varianza di stimatori di ponderazione vincolata complessi si suggerisce di leggere i *paragrafi 1.3.1.1, 1.3.1.2, 1.3.1.3 e 1.3.1.4.*

❑ **Stimatore di Horvitz-Thompson ed espansione**

Nel caso in cui si voglia calcolare la varianza di uno stimatore Horvitz-Thompson o espansione, è necessario che:

- la variabile POP_PIAN assuma un valore costante su tutto il *data-set*;
- la variabile COEF_FIN sia uguale alla variabile COEF_DIR;
- sia presente un'unica variabile X_1 , ottenuta come: $X_1 = 1/\text{COEF_DIR}$;
- sia $CK = 1/\text{COEF_DIR}$;
- la variabile STRATO sia costante se il disegno non è stratificato, mentre se il disegno è stratificato il valore di tale variabile indichi lo strato al quale appartiene il record.

Una forma alternativa del *data-set* di input che tiene conto di pesi costruiti con lo stimatore di Horvitz-Thompson o espansione prevede che:

- la variabile POP_PIAN assuma un valore costante su tutto il data set;
- la variabile COEF_FIN sia uguale alla variabile COEF_DIR;
- non siano presenti variabili X_j ;
- sia $CK = 1$;
- la variabile STRATO sia costante se il disegno non è stratificato, mentre se il disegno è stratificato il valore di tale variabile indichi lo strato al quale appartiene il record.

Infine è utile ricordare che per i disegni semplici, quando il campione non presenta mancate risposte totali, lo stimatore espansione deve avere la variabile COEF_DIR pari a N/n , in cui N è la numerosità della popolazione obiettivo mentre n è la dimensione del campione.

E' importante ricordare che questo tipo di stimatore non richiede una fase di calibrazione dei coefficienti diretti e quindi non bisogna utilizzare la funzione *Riponderazione*.

❑ **Stimatore rapporto**

Di seguito sono illustrate le caratteristiche dei *data-set* di input necessarie per ottenere la stima della varianza dei principali stimatori del rapporto.

- ❑ Stimatore di Hájek (variabile ausiliaria: numerosità di popolazione)
 - la variabile POP_PIAN assume un valore costante su tutto il *data-set*;
 - si presenta un'unica variabile ausiliaria $X_1=1$;
 - si pone $CK=1$;
 - la variabile STRATO è costante.
- ❑ Stimatore del rapporto semplice
 - la variabile POP_PIAN assume un valore costante su tutto il *data-set*;
 - si presenta un'unica variabile ausiliaria X_1 che assume i valori osservati sulle unità campionarie della variabile ausiliaria x su cui si basa lo stimatore;
 - si pone $CK=X_1$;
 - la variabile STRATO è costante.

❑ **Stimatore del rapporto separato**

La stima della varianza si può ottenere definendo il *data-set* INP secondo diverse alternative⁶ che dipendono dalla definizione congiunta delle variabili POP_PIAN e delle variabili X_j nel *data-set* INP. Bisogna inoltre distinguere il caso in cui la stratificazione è ottenuta con una variabile semplice o è il risultato di una classificazione incrociata di più variabili.

Nel primo caso il *data-set* di input può essere costruito facendo riferimento a due schemi alternativi, denominati schema A e B (per approfondimenti cfr. *paragrafo 1.3.1.1*).

⁶

Questo tipo di stimatore prevede la stratificazione della popolazione obiettivo e ciascuno strato rappresenta un gruppo di riferimento del modello secondo la terminologia degli stimatori di regressione generalizzata e di ponderazione vincolata. Nel paragrafo 1.3.1.1, sono approfonditamente illustrate le possibili alternative per definire i gruppi di riferimento del modello.

Seguendo lo schema A si ha che:

- la variabile POP_PIAN è costante;
- si presentano tante variabili ausiliarie X_j per quanti sono gli strati del disegno;
- ogni variabile X_j è associata ad uno strato;
- per ogni record, tutte le variabili X_j sono nulle tranne quella associata allo strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria X su cui si basa lo stimatore (nel caso dello stimatore separato di Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile STRATO indica lo strato al quale appartiene il record.

Per quanto riguarda lo schema B,

- la variabile POP_PIAN assume un valore costante su tutte le unità appartenenti ad uno strato. Le unità appartenenti a strati diversi presentano valori diversi della variabile POP_PIAN;
- si presenta un'unica variabile X_1 che assume i valori osservati sulle unità campionarie dalla variabile ausiliaria x su cui si basa lo stimatore (nel caso dello stimatore separato di Hájek la variabile assume valore “1”);
- si pone $CK=X_1$;
- la variabile STRATO indica lo strato al quale appartiene il record.

Quando la variabile di stratificazione è ottenuta dalla combinazione delle modalità di più variabili (variabile combinata) è possibile anche definire il *data-set* tenendo in considerazione la costruzione del gruppo di riferimento secondo uno schema alternativo, denominato schema C (per approfondimenti cfr. *paragrafo 1.3.1.1*). In questo caso, le variabili originali di stratificazione sono divise in due gruppi complementari. Il primo gruppo definisce il numero di valori che assume la variabile POP_PIAN, attraverso la combinazione delle modalità delle variabili ad esso appartenenti; mentre il secondo gruppo determina, mediante la classificazione incrociata delle modalità delle variabili in esso contenute, il numero di variabili ausiliarie X_j . Si ha quindi che:

- la variabile POP_PIAN assume tanti valori per quante sono le com-

binazioni delle modalità del primo gruppo di variabili di stratificazione. Ad ogni combinazione di modalità è associato un valore diverso della variabile POP_PIAN;

- per ciascun record il valore della variabile POP_PIAN è quello associato alla combinazione delle modalità osservate sul record stesso;
- si presentano tante variabili ausiliarie X_j pari al numero delle combinazioni delle modalità del secondo gruppo di variabili di stratificazione;
- per ogni record tutte le variabili X_j sono nulle tranne quella associata alla combinazione di modalità che presenta il record. Questa variabile assume il valore della variabile ausiliaria X su cui si basa lo stimatore (nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile STRATO indica lo strato al quale appartiene il record.

La scelta di una tra le diverse alternative deve essere dettata dai vincoli operativi (presenti nello schema C) e dalla struttura del *data-set* INP prima che venga modificato secondo uno degli schemi. Altri suggerimenti sono forniti nel *paragrafo 1.3.1.2*.

□ Stimatore del rapporto combinato

Per questo tipo di stimatore, il *data-set* deve presentare i seguenti requisiti:

- la variabile POP_PIAN assume un valore costante su tutto il *data-set*;
- si presenta un'unica variabile ausiliaria X_1 che assume i valori osservati della variabile ausiliaria X su cui si basa lo stimatore;
- si pone $CK=X_1$;
- la variabile STRATO indica lo strato al quale appartiene il record.

□ Stimatore rapporto post-stratificato

Per tenere nella giusta considerazione la costruzione dei coefficienti finali ottenuti con questo stimatore il *data-set* di input deve presentare le seguenti caratteristiche:

- la variabile POP_PIAN è costante;
- si presentano tante variabili X_j per quanti sono i post-strati del disegno;
- ogni variabile X_j è associata ad un post-strato;
- per ogni record tutte le variabili X_j sono nulle tranne quella associata al post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile STRATO è costante.

□ Stimatore rapporto post-stratificato separato

Per questo tipo di stimatore il *data-set* di input può presentare tre strutture⁷. La prima, detta schema A, prevede che:

- la variabile POP_PIAN sia costante;
- siano presenti tante variabili X_j per quante sono le combinazioni tra le modalità degli strati e dei post-strati del disegno;
- ogni variabile X_j sia associata ad una combinazione tra uno strato e un post-strato;
- per ogni record tutte le variabili X_j siano nulle tranne quella associata alla combinazione strato per post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile STRATO indica lo strato al quale appartiene il record.

Seguendo una seconda impostazione, detta schema B, il *data-set* presenta le seguenti caratteristiche:

- la variabile POP_PIAN assume tanti valori per quante sono le

⁷

Questo tipo di stimatore prevede che il post-strato all'interno dello strato rappresenti un gruppo di riferimento del modello secondo la terminologia degli stimatori di regressione generalizzata e di ponderazione vincolata. Nel paragrafo 1.3.1.1, sono approfonditamente illustrate le possibili alternative per definire i gruppi di riferimento del modello.

modalità della variabile di stratificazione (numero di strati). Ad ogni strato è associato un valore diverso della variabile POP_PIAN;

- per ciascun record il valore della variabile POP_PIAN è quello associato allo strato in cui si trova il record stesso;
- si presentano tante variabili ausiliarie X_j pari al numero dei post-strati;
- per ogni record tutte le variabili X_j sono nulle, tranne quella associata al post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria su cui si basa lo stimatore (nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile STRATO indica lo strato al quale appartiene il record.

Nel caso in cui la stratificazione avviene secondo due o più variabili, la variabile POP_PIAN e le variabili X_j possono essere specificate secondo un terzo schema alternativo, detto schema C. Nella seguente tabella è descritta sinteticamente l'impostazione del *data-set* di input basata sugli schemi A, B e C.

<u>POP_PIAN è identificata da:</u> Nessuna variabile di stratificazione (schema A)	⇒	<u>$X_1, \dots, X_j, \dots, X_J$ sono identificate da:</u> Le modalità ottenute dalla combinazione tra la variabile di post-stratificazione e le variabili di stratificazione
Una variabile di stratificazione (schema C)	⇒	Le modalità ottenute dalla combinazione tra la variabile di post-stratificazione e le restanti variabili di stratificazione
La combinazione di un sottoinsieme di variabili di stratificazione (schema C)	⇒	Le modalità ottenute dalla combinazione tra la variabile di post-stratificazione e le restanti variabili di stratificazione
La combinazione di tutte le variabili di stratificazione (schema B)	⇒	Le modalità della variabile di post-stratificazione

La scelta di una tra le diverse alternative deve essere dettata dai vincoli operativi (presenti nello schema C) e dalla struttura del *data-set* INP prima che venga modificato secondo uno degli schemi. Altri suggerimenti sono forniti nel *paragrafo 1.3.1.2*.

□ Stimatore rapporto post-stratificato combinato

Per tale stimatore il *data-set* di input presenta la seguente forma:

- la variabile POP_PIAN è costante;

- si presentano tante variabili X_j per quanti sono i post-strati del disegno;
- ogni variabile X_j è associata ad un post-strato;
- per ogni record tutte le variabili X_j sono nulle, tranne quella associata al post-strato in cui si trova il record. Questa variabile assume il valore della variabile ausiliaria X su cui si basa lo stimatore (nel caso dello stimatore separato tipo Hájek la variabile assume valore “1”);
- si pone CK pari al valore della variabile X_j non nulla;
- la variabile $STRATO$ indica lo strato al quale appartiene il record.

□ **Stimatore raking**

I $COEF_FIN$ ottenuti dagli stimatori raking sono calibrati sui totali di popolazione di una sola variabile ausiliaria per sottopopolazioni, denominate gruppi di riferimento, appartenenti a diverse partizioni distinte della popolazione obiettivo (cfr. *appendice A.1*). In particolare lo stimatore ratio raking considera due partizioni in gruppi di riferimento, mentre lo stimatore raking generalizzato estende la calibrazione a totali per più di due partizioni in gruppi di riferimento della popolazione obiettivo.

□ **Stimatore ratio raking**

- Siano Q_1 e Q_2 il numero di gruppi di riferimento di una popolazione, definiti rispettivamente sulla base delle modalità assunte dalle variabili ausiliarie v_1 e v_2 .

Per definire il *data-set* INP occorre che:

- la variabile POP_PIAN assume un valore costante su tutto il *data-set*;
- a ciascuno dei Q_1+Q_2 gruppi di riferimento si associa una variabile X_j con $j=1, \dots, Q_1+Q_2$ ⁸;
- ciascuna variabile X_j assume valore nullo, tranne quella che corrisponde al gruppo di riferimento a cui appartiene il record. In que-

⁸

Si ricorda che i nomi X_j assegnati alle variabili per definire i gruppi di riferimento sono utilizzati con un puro scopo descrittivo. In realtà non esistono vincoli particolari sul tipo di nome da assegnare, e quindi, a maggior ragione, non è necessario attribuire ai nomi delle Variabili ausiliarie un indice.

sto caso X_j è posta pari a “1” (per ogni record sono presenti due variabili X_j pari a “1”);

- la variabile $CK=1$ per tutti i record del *data-set*.

□ Stimatore raking generalizzato

L'impostazione dei *data-set* di input per ottenere i coefficienti di riporto finali dello stimatore è facilmente ricavabile da quanto illustrato per lo stimatore ratio raking. Questo stimatore generalizza il precedente, considerando un insieme di variabili qualitative (V_1, \dots, V_G) che definiscono $G(>2)$ partizioni in gruppi di riferimento.

Riprendendo la simbologia e le regole descritte per lo stimatore ratio raking è, pertanto, necessario definire $Q_1 + \dots + Q_G$ variabili X_j .

La costruzione dei *data-set* avviene in analogia a quanto descritto per lo stimatore ratio raking.

□ Stimatori di regressione generalizzata e di ponderazione vincolata

Gli stimatori di regressione generalizzata o la più ampia classe degli stimatori di ponderazione vincolata utilizzano i totali noti per sottopopolazioni, denominate gruppi di riferimento, appartenenti a partizioni distinte della popolazione obiettivo (cfr. *appendice A.1*). Per tale struttura dei totali noti la costruzione dei *data-set* di input può seguire diverse alternative che sono descritte nei *paragrafi 1.3.1.1 e 1.3.1.2*.

Un esempio dettagliato e molto generale che descrive la costruzione dei due *data-set* di input è illustrato nell'*appendice A.4*.

□ Stimatori a livello di cluster

Per definire gli stimatori visti in precedenza a livello di cluster, quando i record del *data-set* si riferiscono alle unità elementari, è sufficiente operare sulla sola variabile CK . Più precisamente, il valore della variabile CK deve essere così attribuito:

- CK è uguale alla somma dei valori che tale variabile assume sui

record che appartengono al cluster quando si definisce il modello *a livello di unità elementare*.

A titolo di esempio, indicato con M il numero di record inclusi in un generico cluster, si ha per lo stimatore di Hájek, $CK=M$.

1.3.2 Definizione delle variabili di input per un dato disegno

Sintesi: Il paragrafo descrive la costruzione del data-set di input in funzione del disegno campionario adottato dall'utente. In particolare si illustrano i criteri di costruzione del data-set di input per il:

- campionamento stratificato di unità elementari con reimmissione e probabilità di selezione costanti (1.3.2.1)
- campionamento stratificato di grappoli di unità elementari con reimmissione e probabilità di selezione costanti (1.3.2.2)
- campionamento stratificato di unità elementari senza reimmissione e probabilità di inclusione costante (1.3.2.3)
- campionamento stratificato di grappoli di unità elementari senza reimmissione e probabilità di inclusione costante (1.3.2.4)
- campionamento stratificato di unità elementari con o senza reimmissione e probabilità di inclusione variabile (1.3.2.5)
- campionamento stratificato di grappoli di unità elementari con o senza reimmissione e probabilità di inclusione variabile (1.3.2.6)
- campionamento a due o più stadi di selezione (1.3.2.7).

Per i principali disegni campionari, il software dispone di uno stimatore corretto o approssimativamente corretto della varianza campionaria. Per selezionare lo stimatore della varianza legato al disegno di campionamento che ha dato origine ai coefficienti finali di riporto (COEF_FIN) l'utente deve agire su alcune variabili del *data-set* di input, denominate TIPO_DIS, STRATO, UNITA_2, UNITA_1 (cfr. tabella 1.2). I valori attribuiti a queste ultime indicano al software il tipo di formula per la stima della varianza che deve essere utilizzata.

Il paragrafo illustra i requisiti essenziali delle variabili per i disegni di campionamento stratificati, tralasciando il caso dei disegni semplici. Per questi ultimi piani campionari, l'unica differenza risiede nella definizione della variabile STRATO. Mentre nei disegni stratificati, come si vedrà in seguito, la variabile STRATO assume differenti valori (pari al numero

degli strati del piano di campionamento), nei disegni semplici tale variabile di input è costante su tutti i record.

Infine, è opportuno rilevare che il software consente di stimare la varianza anche per campioni le cui unità sono state estratte con disegni di campionamento differenti. Per chiarire tale aspetto si consideri l'esempio 1.3.

Esempio 1.3:

Nelle principali indagini sulle famiglie condotte dall'ISTAT, l'universo è essenzialmente suddiviso in due sottopopolazioni; la prima è costituita dalle famiglie residenti nei Comuni di "piccole dimensioni demografiche", la seconda è rappresentata dalle famiglie residenti nei Comuni di "grandi dimensioni demografiche" (incluso in alcune indagini i Comuni capoluogo di regione). Per le due sottopopolazioni il campione di famiglie viene estratto secondo due disegni distinti: per la prima sottopopolazione il disegno è a due stadi di campionamento in cui le unità di primo stadio sono i Comuni e le unità di secondo stadio sono le famiglie; per la seconda sottopopolazione il disegno prevede per ogni Comune la selezione senza reimmissione con probabilità costante delle famiglie. Tutti i componenti delle famiglie estratte vengono intervistati.

Con questo schema di campionamento composto, i record relativi ai singoli componenti delle famiglie del campione appartenenti ai Comuni di "piccole dimensioni demografiche" assumono i valori nelle variabili TIPO_DIS, STRATO, UNITA_2, UNITA_1 secondo i punti illustrati nel paragrafo 13.2.7. Per i record che individuano le unità appartenenti alla seconda sottopopolazione, le variabili sono definite secondo quanto è descritto nel paragrafo 1.3.2.4.

1.3.2.1 CAMPIONAMENTO STRATIFICATO DI UNITÀ ELEMENTARI CON REIMMISSIONE E CON PROBABILITÀ DI SELEZIONE COSTANTE

Per utilizzare lo stimatore corretto o asintoticamente corretto della varianza campionaria implementato dal software, le variabili del *data-set* di input TIPO_DIS, STRATO, UNITA_2, UNITA_1 devono presentare le seguenti caratteristiche:

- la variabile TIPO_DIS assume valori pari a "0" su tutti i record;
- la variabile STRATO deve assumere tanti valori distinti per quanti sono gli strati;
- la variabile STRATO deve assumere un valore uguale per tutti i

record del *data-set* appartenenti allo stesso strato; record appartenenti a strati diversi presentano valori diversi della variabile STRATO (nel caso di un campione non stratificato, la variabile STRATO deve assumere un valore costante per tutte le unità del data set);

- per ciascun record la variabile UNITA_2 deve assumere un valore univoco su tutto il *data-set*;
- se un'unità elementare è stata selezionata due volte (o più), quest'ultima deve apparire nel *data-set* di input come due (o più) record distinti, aventi cioè due (o più) valori diversi della variabile UNITA_2;
- la variabile UNITA_1 assume i medesimi valori della variabile UNITA_2;
- i record con lo stesso valore della variabile UNITA_2 devono presentare anche lo stesso valore della variabile STRATO.

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di UNITA_1 quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.2 CAMPIONAMENTO STRATIFICATO DI GRAPPOLI DI UNITÀ ELEMENTARI CON REIMMISSIONE E PROBABILITÀ DI SELEZIONE COSTANTE

E' necessario ricordare che per questo tipo di disegno ciascun record del *data-set* di input rappresenta un'unità elementare appartenente ad un grappolo. Affinché il software utilizzi lo stimatore corretto o approssimativamente corretto per questo tipo di disegno è necessario agire nel seguente modo sulle variabili del *data-set*:

- si pone la variabile TIPO_DIS pari a "0" su tutti i record;
- la variabile STRATO deve assumere tanti valori distinti quanti sono gli strati;
- la variabile STRATO deve assumere un valore uguale per tutti i record appartenenti a grappoli contenuti nello stesso strato; record appartenenti a grappoli contenuti in strati diversi presentano valori diversi della variabile STRATO;

- la variabile `UNITA_2` assume tanti valori distinti pari al numero di grappoli presenti nel *data-set*. A ciascun valore della variabile `UNITA_2` è associato un grappolo di unità elementari e quindi i record appartenenti allo stesso grappolo presentano lo stesso valore della variabile `UNITA_2`; record appartenenti a grappoli diversi devono avere un diverso valore della variabile `UNITA_2`;
- se un grappolo è stato selezionato due (o più) volte nel campione, quest'ultimo deve apparire nel *data-set* come due (o più) grappoli distinti, aventi cioè due (o più) valori diversi della variabile `UNITA_2`; pertanto ogni record appartenente ad un grappolo selezionato due (o più) volte nel campione, deve apparire nel *data-set* di input come due (o più) record distinti, aventi cioè due (o più) valori diversi della variabile `UNITA_2`;
- la variabile `UNITA_1` assume i medesimi valori della variabile `UNITA_2`;
- i record con lo stesso valore della variabile `UNITA_2` devono presentare anche lo stesso valore della variabile `STRATO`.

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di `UNITA_1` quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.3 CAMPIONAMENTO STRATIFICATO DI UNITÀ ELEMENTARI SENZA REIMMISSIONE E PROBABILITÀ DI INCLUSIONE COSTANTE

Se i coefficienti di riporto all'universo provengono da una strategia campionaria che prevede questo tipo di disegno, per adottare lo stimatore della varianza campionaria corretto o approssimativamente corretto, le variabili del *data-set* devono essere così definite:

- si pone la variabile `TIPO_DIS` pari a "1" su tutti i record;
- relativamente alle variabili `STRATO`, `UNITA_2`, `UNITA_1` si effettuano le medesime operazioni indicate per il disegno campionario di unità elementari con reimmissione e probabilità di selezione costante (*paragrafo 1.3.2.1*).

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di UNITA_1 quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.4 CAMPIONAMENTO STRATIFICATO DI GRAPPOLI DI UNITÀ ELEMENTARI SENZA REIMMISSIONE E PROBABILITÀ DI INCLUSIONE COSTANTE

Dal punto di vista operativo, per scegliere lo stimatore della varianza campionaria corretto o approssimativamente corretto, implementato dal software, si agisce nel modo seguente sulle variabili del disegno:

- si pone la variabile TIPO_DIS pari a “1” su tutti i record;
- relativamente alle variabili STRATO, UNITA_2, UNITA_1 si effettuano le medesime operazioni indicate nell’analogo disegno che prevede la selezione con reimmissione dei grappoli e probabilità di selezione costanti (*paragrafo 1.3.2.2*).

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di UNITA_1 quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.5 CAMPIONAMENTO STRATIFICATO DI UNITÀ ELEMENTARI CON O SENZA REIMMISSIONE E CON PROBABILITÀ DI INCLUSIONE VARIABILE

I due disegni di campionamento, con e senza reimmissione, si trattano in modo congiunto in quanto il software implementa solo lo stimatore corretto della varianza per piani in cui la selezione è con reimmissione. Tale stimatore è invece distorto positivamente per i disegni senza reimmissione. Tuttavia, quando il tasso di campionamento delle unità all’interno degli strati è “piccolo”, questo diventa approssimativamente corretto. Per implementare tale stimatore le variabili del *data-set* di input presentano le seguenti proprietà:

- la variabile TIPO_DIS è pari a “0” su tutti i record;
- la variabile STRATO deve assumere tanti valori distinti quanti sono gli strati;

- la variabile STRATO deve assumere un valore uguale per tutti i record del *data-set* appartenenti allo stesso strato; record appartenenti a strati diversi presentano valori diversi della variabile STRATO;
- per ciascun record la variabile UNITA_2 deve assumere un valore univoco su tutto il data set;
- nel caso di selezione con reimmissione, se un'unità elementare è stata selezionata due volte (o più), quest'ultima deve apparire nel *data-set* di input come due (o più) record distinti, aventi cioè due (o più) valori diversi della variabile UNITA_2;
- la variabile UNITA_1 assume i medesimi valori della variabile UNITA_2;
- i record con lo stesso valore della variabile UNITA_2 devono presentare anche lo stesso valore della variabile STRATO.

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di UNITA_1 quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.6 CAMPIONAMENTO STRATIFICATO DI GRAPPOLI DI UNITÀ ELEMENTARI CON O SENZA REIMMISSIONE E PROBABILITÀ DI INCLUSIONE VARIABILI

Anche in questo caso, sia per il campionamento con reimmissione che per quello senza reimmissione, la varianza viene calcolata utilizzando il metodo adottato per il campionamento con reimmissione. In analogia con quanto visto per il disegno a grappoli con probabilità di selezione costanti, le variabili del *data-set* devono essere definite nel seguente modo:

- si pone la variabile TIPO_DIS pari a "0" su tutti i record;
- la variabile STRATO deve assumere tanti valori distinti quanti sono gli strati;
- la variabile STRATO deve assumere un valore uguale per tutti i record appartenenti a grappoli contenuti nello stesso strato; record appartenenti a grappoli contenuti in strati diversi presentano valori diversi della variabile STRATO;
- la variabile UNITA_2 assume tanti valori distinti pari al numero di

grappoli presenti nel data set. A ciascun valore della variabile `UNITA_2` è associato un grappolo di unità elementari e quindi i record appartenenti allo stesso grappolo presentano lo stesso valore della variabile `UNITA_2`; record appartenenti a grappoli diversi devono avere un diverso valore della variabile `UNITA_2`;

- nel caso di selezione con reimmissione, se un grappolo è stato selezionato due (o più) volte nel campione, quest'ultimo deve apparire nel *data-set* come due (o più) grappoli distinti, aventi cioè due (o più) valori diversi della variabile `UNITA_2`; pertanto ogni record appartenente ad un grappolo selezionato due (o più) volte nel campione, deve apparire nel *data-set* di input come due (o più) record distinti, aventi cioè due (o più) valori diversi della variabile `UNITA_2`;
- la variabile `UNITA_1` assume i medesimi valori della variabile `UNITA_2`;
- i record con lo stesso valore della variabile `UNITA_2` devono presentare anche lo stesso valore della variabile `STRATO`.

Nel *paragrafo 1.3.2.8* vengono evidenziate alcune considerazioni sulla procedura che definisce correttamente i valori di `UNITA_1` quando il disegno è composto da un campionamento ad uno stadio e un campionamento a due stadi.

1.3.2.7 CAMPIONAMENTO A DUE O PIÙ STADI DI SELEZIONE

Tra i diversi stimatori della varianza campionaria, il software implementa quello corretto (o asintoticamente corretto per stimatori linearizzabili in serie di Taylor) per i disegni a due o più stadi con reimmissione delle unità di primo stadio. Il software non prevede il calcolo di uno stimatore corretto della varianza quando i coefficienti di riporto all'universo provengono da un disegno a due o più stadi di campionamento senza reimmissione delle unità primarie. Quindi, anche in quest'ultimo caso, l'utente deve scegliere lo stimatore della varianza per disegni con reimmissione, il quale, tuttavia, risulta distorto per i disegni senza reimmissione. Tale distorsione è, comunque, trascurabile quando il tasso di campionamento delle unità primarie all'interno degli strati è "piccolo".

Inoltre, poiché la forma funzionale dello stimatore non cambia a secon-

da che la probabilità di selezione sia costante o variabile, non è necessario trattare separatamente le due diverse strategie di estrazione delle unità primarie. Pertanto, per i disegni a due o più stadi di campionamento lo stimatore della varianza campionaria corretto o approssimativamente corretto si richiama con gli stessi dati di input. In particolare:

- si pone la variabile TIPO_DIS pari a “0” su tutti i record;
- la variabile STRATO deve assumere tanti valori distinti quanti sono gli strati;
- la variabile STRATO deve assumere un valore uguale per tutti i record appartenenti a unità primarie contenute nello stesso strato; record appartenenti a unità primarie contenute in strati diversi presentano valori diversi della variabile STRATO;
- la variabile UNITA_1 assume tanti valori distinti pari al numero di unità primarie contenute nel *data-set*. A ciascun valore della variabile UNITA_1 è associata un’unità primaria e quindi all’interno di uno strato i record appartenenti alla stessa unità primaria presentano lo stesso valore della variabile UNITA_1; record appartenenti a unità primarie diverse devono avere un diverso valore della variabile UNITA_1;
- nel caso di selezione con reimmissione, se un’unità primaria è stata selezionata due (o più) volte nel campione, quest’ultima deve apparire nel *data-set* come due (o più) unità distinte aventi, cioè, due (o più) valori diversi della variabile UNITA_1; pertanto ogni record appartenente ad un’unità primaria selezionata due (o più) volte nel campione, deve apparire nel *data-set* di input come due (o più) record distinti, aventi cioè due (o più) valori diversi della variabile UNITA_1;
- la variabile UNITA_2 assume tanti valori distinti pari al numero di unità finali. A ciascun valore della variabile UNITA_2 è associata un’unità finale. Per i disegni a due stadi l’unità finale coincide con quella di secondo stadio. Pertanto se le unità di secondo stadio sono grappoli di unità elementari la variabile UNITA_2 è costante per tutti i record appartenenti al grappolo. Se l’unità di secondo stadio è elementare, la variabile UNITA_2 presenta un valore univoco per ogni record. Per i disegni a tre o più stadi di campionamento l’uni-

tà finale a cui è assegnata una modalità identificativa della variabile UNITA_2, è quella unità, classificata all'ultimo stadio di campionamento, oltre il quale non avviene un processo di estrazione casuale di unità elementari. Anche in questo caso l'unità può essere elementare o rappresentare un grappolo di unità elementari. La definizione della variabile UNITA_2 dovrà seguire gli stessi criteri visti per i disegni a due stadi.

1.3.2.8 NOTA SULLA DEFINIZIONE DELLA VARIABILE UNITÀ PRIMARIA (UNITA_1) PER I DISEGNI DI CAMPIONAMENTO COMPOSTI

Per i disegni che si compongono di un campionamento ad uno stadio ed un campionamento a due stadi bisogna fare attenzione ai valori attribuiti alla variabile UNITA_1, poiché in alcuni casi si possono produrre delle duplicazioni errate dei valori di questa variabile. A tale scopo si riprenda il contesto d'indagine dell'esempio 1.3.

Esempio 1.4:

Nel precedente esempio si è detto che in alcune indagini sulle famiglie il campione proveniente dai comuni di piccole dimensioni viene estratto secondo un disegno a due stadi in cui le unità di primo stadio sono rappresentate dai comuni mentre le unità finali e di secondo stadio sono le famiglie.

Per questo disegno la variabile UNITA_1 deve, quindi, identificare univocamente un comune selezionato nel campione e la variabile UNITA_2 deve identificare univocamente la famiglia rilevata.

Il campione di famiglie proveniente dai comuni di grandi dimensioni viene, invece, estratto con un disegno ad uno stadio (tutti i comuni di grandi dimensioni vengono inclusi nel campione) in cui le unità di primo stadio e finali sono le famiglie. Nei record estratti secondo questo disegno la variabile UNITA_1 deve avere gli stessi valori della variabile UNITA_2.

Prima di definire le variabili UNITA_1 e UNITA_2 può accadere che l'utente abbia in partenza nell'archivio di input altre due variabili: la prima che identifica i comuni dell'indagine, denominata in seguito COM, la seconda che identifica le famiglie, denominata in seguito FAM_COD.

L'utente potrebbe quindi costruire la variabile UNITA_1 seguendo queste due rego-

le: UNITA_1 presenta i valori di COM per i record estratti con un disegno a due stadi, ed è pari a FAM_COD per i record estratti con un disegno a uno stadio.

Tuttavia, se viene impiegato questo criterio per la costruzione della variabile è necessario fare attenzione alle eventuali duplicazioni dei valori che si possono creare sulla variabile UNITA_1 e che producono un errore nella procedura di stima della varianza.

Nella tabella seguente è presentato un esempio di tali duplicazioni, dove applicando la procedura sopra descritta si hanno record con lo stesso valore della variabile UNITA_1 ma un diverso valore della variabile STRATO.

Tabella 1.9- Costruzione errata della variabile UNITA_1

Disegno	COM	FAM_COD	UNITA_1	STRATO
UNO STADIO	1	1	1	A
	1	2	2	A
	1	3	3	A
DUE STADI	2	4	2	B
	2	5	2	B
	2	6	2	B

1.4 Definizione delle variabili di input per il livello della stima (dominio di stima) considerato

Sintesi: Nel presente paragrafo si definisce la costruzione di alcune variabili di input sulla base del livello della stima che si desidera ottenere: le stime possono essere calcolate per domini pianificati (paragrafo 1.4.1) e per domini non pianificati (paragrafo 1.4.2).

Il software permette di calcolare la varianza delle stime a livello dell'intera popolazione di riferimento o a livello di una sottopopolazione in essa contenuta, nota in letteratura con il nome di *dominio di stima* (o di *studio*). Per ottenere la stima della varianza a livello di sottopopolazione l'utente deve sapere se i domini di studio che intende esaminare sono di tipo *pianificato* (o *stratificato*) oppure di tipo *non pianificato* (o *non stratificato*).

Un dominio di stima si dice pianificato quando contiene tutte le unità

della popolazione appartenenti ad uno strato oppure ad aggregazioni di strati. Viceversa un dominio non pianificato contiene solo una parte delle unità della popolazione appartenenti ad uno strato del disegno.

Qualora si sia interessati a calcolare le stime delle varianze per un insieme di domini pianificati che formano una partizione dell'universo di riferimento si veda il *paragrafo 1.4.1*; se, invece, si desidera ottenere le stime delle varianze per diversi insiemi di domini pianificati e non pianificati, ciascuno dei quali determina una differente partizione dell'universo si veda il *paragrafo 1.4.2*. In seguito le variabili di input direttamente interessate alla stima della varianza per domini sono richiamate attraverso specifiche denominazioni (cfr. tabella 1.10)

Tabella 1.10 - Variabili del data-set di input per definire il livello delle stime da analizzare

Variabili di input (paragrafo 1.1)	Nome sintetico della variabile
Variabili di sottoclasse	S1, ..., Sc, ..., SC (ogni variabile indica una partizione in domini di stima dell'universo di riferimento)
Dominio pianificato	DOMSTIMA

1.4.1 Definizione delle variabili di input per i domini di stima pianificati

Quando tutti i domini di stima che si vogliono esaminare formano una partizione dell'universo, secondo domini di studio pianificati, l'utente deve intervenire sulla definizione della variabile DOMSTIMA. Premesso che ad ogni dominio di stima è stato assegnato un codice identificativo, la variabile di input deve presentare le seguenti caratteristiche:

- per ciascun record la variabile DOMSTIMA presenta un codice che è quello associato al dominio di stima al quale appartiene il record stesso;
- la variabile DOMSTIMA assume un unico codice su tutte le unità appartenenti ad uno stesso strato. Inoltre, le unità appartenenti a strati contenuti nello stesso dominio di stima presentano lo stesso codice di DOMSTIMA. Le unità appartenenti a strati contenuti in diversi domini di stima devono presentare un codice diverso in DOMSTIMA;

- nel caso in cui il dominio di stima è rappresentato dall'intero universo di riferimento la variabile DOMSTIMA presenta un unico codice sul tutto il *data-set* di input.

E' necessario ricordare che è obbligatorio inserire la variabile DOMSTIMA nell'archivio di input anche quando non si richiede una stima della varianza per domini di studio pianificati. In questa circostanza si può attribuire un codice costante alla variabile su tutto il data set. Il software in tal caso fornisce le stime riferite a tutta la popolazione obiettivo.

Infine per ciascun dominio di stima, il software offre come output anche una rappresentazione sintetica degli errori di campionamento (cfr. *appendice A.5*).

1.4.2 Definizione delle variabili di input per i domini di stima non pianificati

Per richiedere al software il calcolo della stima della varianza per una serie di domini non pianificati, che nel loro complesso formano una partizione dell'universo di riferimento, l'utente deve costruire la variabile S1, detta variabile di *sottoclasse*, le cui modalità identificano gli specifici domini in questione. La variabile S1 presenta la seguente caratteristica:

- per ciascun record la variabile S1 presenta il codice che identifica il dominio di stima non pianificato al quale appartiene il record stesso.

Il software permette di considerare contemporaneamente diverse partizioni di domini non pianificati della popolazione di riferimento.

Operativamente la procedura richiede l'inserimento di tante variabili di input, per quante sono le partizioni che si vogliono considerare. Se ad esempio si vogliono stimare C partizioni si inseriranno nell'archivio di input C variabili di sottoclasse denominate S1, ..., Sc, ..., SC.

Ogni variabile è costruita, come descritto nel punto precedente, con riferimento ai domini di stima di una specifica partizione.

E' utile osservare che la generica variabile Sc ($c=1, \dots, C$) può anche identificare una partizione costituita da domini pianificati. Tuttavia, il software è più efficiente quando si stimano le varianze utilizzando la variabile

DOMSTIMA. Inoltre, ricordando che la variabile DOMSTIMA deve essere definita obbligatoriamente nel *data-set* di input, il software produce la stima della varianza per:

- i domini pianificati identificati dalle modalità della variabile DOMSTIMA;
- i domini non pianificati identificati dalle modalità della variabile Sc;
- i domini non pianificati identificati dalla combinazione delle modalità della variabile DOMSTIMA e Sc.

Infine, è importante sottolineare che per i domini non pianificati il software non prevede una rappresentazione sintetica degli errori campionari.

2. I data-set di output

Sintesi: Il software produce alcuni data-set di output scritti sulla cartella di output scelta dall'utente.

I data-set di output sono i seguenti⁹:

- Data-set creato per memorizzare parametri di input (paragrafo 2.1)
SAVEPAR
- Data-set creato per memorizzare gli errori rilevati sull'input (paragrafo 2.2)
ERRORI_INPUT
- Data-set contenenti le informazioni relative a stime ed errori campionari (paragrafo 2.2)
STRATO, TOTALE, TOT_DIS0, TOT_DIS1
- Data-set contenenti informazioni sulla stratificazione e sul campione (paragrafo 2.4)
TAB1, UNIC
- Data-set contenenti le informazioni relative a stime ed errori campionari utili ad elaborazioni successive (paragrafo 2.5)
WSTRATO, WTOTALE, WTOT_DIS0, WTOT_DIS1

2.1 Il data-set dei parametri di input

Nel *paragrafo 5.2.3 della Sezione I* viene descritto come selezionare le variabili di input tramite i parametri della procedura. Ciò è possibile in quanto

⁹ La cartella di output scelta dall'utente corrisponde alla libreria "errori". Se, ad esempio, l'utente sceglie la cartella c:\utente - prendendo in considerazione il data-set di output STRATO - la procedura crea il data-set Sas di output "errori.strato" che corrisponde al file c:\utente\STRATO.sas7bdat (data-set sas v.8) registrato nella cartella c:\utente. Per semplificare l'esposizione successiva si farà riferimento ai data-set solo con il nome, senza l'estensione del file o la libreria di riferimento.

il software, per ciascuna elaborazione, scrive nella cartella di output il *data-set* SAVEPAR, indispensabile per attivare la funzione “Parametri attivi”.

In figura 2.1 è possibile vedere un esempio di *data-set* SAVEPAR: il *data-set* è caratterizzato da due soli campi “*descr*” e “*parametro*” e il software scrive in automatico le sedici righe del *data-set*, memorizzando le scelte fatte dall’utente per attivare l’applicazione.

Figura 2.1: Il data-set SAVEPAR

	descr	parametro
1	INPUT	C:\applicazioneLiberr.Esempio
2	Variabili di interesse	Y1 Y2 Y3 Y4 Y5 Y6
3	Peso distanza	PESO_DIST
4	Variabili ausiliarie	38 X39 X40 X41 X42 X43 X44 X45 X46 X47 X48
5	Dominio pianificato	PROVINCIA
6	Peso finale	PESO_FIN
7	Sottoclassi	SEX
8	Tipo di disegno	TIPO_DISE
9	Unità primaria	UN_PRIM
10	Unità finale	COD_FAM
11	Strato	STRATO
12	Peso diretto	PESO_INIZ
13	Popolaz.pianif.util. per stimatore	REGIONE
14	Nr. unità superstrato	2
15	Variabili: 1 quant., 2 qual.	2
16	Stima: 1 unità elem., 2 cluster	1

2.2 Gli errori rilevati sul data-set di input

Il software Genesees è predisposto al controllo automatico di alcuni errori rilevati sul *data-set* di input; in tal caso la procedura ferma l’elaborazione e scrive un *data-set* contenente l’informazione relativa. Il *data-set* contenente gli errori è sempre ERRORI_INPUT; l’informazione che viene scritta dipende dall’errore accertato:

- Il software verifica che non vi siano incoerenze riscontrabili tra il codice della variabile di disegno “Unità primaria” e il codice della variabile “Unità finale”. Come visto nel *paragrafo 1.2* non devono esistere unità elementari identificate da diversi codici di “Unità primaria” e da uno stesso codice di “Unità finale”. In caso di errore,

prima che l'elaborazione sia fermata, viene scritto il *data-set* ERRORI_INPUT. Tale *data-set* è caratterizzato da due campi: il primo contiene il codice della “Unità primaria” coinvolto nell'incoerenza riscontrata, il secondo contiene il codice della “Unità finale”: saranno riscontrabili più righe con diversi codici di “Unità primaria” e lo stesso codice di “Unità finale” (cfr. Schema1 – *paragrafo 1.2*).

- Il software verifica che non vi siano incoerenze riscontrabili tra il codice della variabile di disegno “Strato” e il codice della variabile “Unità primaria”. Nel *paragrafo 1.2* è evidenziato che non devono esistere unità elementari identificate da diversi codici della variabile “Strato” e dallo stesso codice di “Unità primaria”. In caso di errore, prima che l'elaborazione sia fermata, viene scritto il *data-set* ERRORI_INPUT. Tale *data-set* è caratterizzato dai due campi in cui vengono riportati i codici delle due variabili di cui sopra. In tale *data-set* dovranno dunque essere riscontrabili più righe con diversi codici di “Strato” e lo stesso codice di “Unità primaria” (cfr. Schema2 – *paragrafo 1.2*).
- La partizione definita dalla variabile relativa allo stimatore “Popolazione pianificata utilizzata per lo stimatore” risulta sempre formata da sottogruppi corrispondenti ad aggregazioni di strati. Uno strato può appartenere ad un solo sottoinsieme della partizione. Per rispettare i vincoli definiti nel *paragrafo 1.2* sulla relazione tra questa variabile e la variabile “Strato” la procedura effettua un controllo e in caso di errore, prima che l'elaborazione sia fermata, viene scritto il *data-set* ERRORI_INPUT. Tale *data-set* è caratterizzato dai due campi in cui vengono riportati i codici delle due variabili di cui sopra. Nel *data-set* dovranno dunque essere riscontrabili più righe con diversi codici della variabile “Popolazioni pianificate utilizzate per lo stimatore” e lo stesso codice della variabile “Strato” (cfr. Schema3 – *paragrafo 1.2*).
- La partizione definita dalla variabile relativa al dominio di stima “Dominio pianificato” risulta sempre formata da sottogruppi corrispondenti ad aggregazioni di strati. Uno strato può appartenere ad un solo sottoinsieme della partizione. Per rispettare i vincoli definiti nel *paragrafo 1.2* sulla relazione tra questa variabile e la variabile

“Strato” la procedura effettua un controllo e nel caso di errore, prima che l’elaborazione sia fermata, scrive il *data-set* ERRORI_INPUT. Tale *data-set* è caratterizzato dai due campi in cui vengono riportati i codici delle due variabili di cui sopra. Nel *data-set* dovranno dunque essere riscontrabili più righe con diversi codici della variabile “Dominio pianificato” e lo stesso codice della variabile “Strato”.

- Le unità elementari appartenenti ad un certo strato avranno tutte il codice della variabile “Tipo di disegno” pari ad “1” o a “0”. In caso di errore, prima che l’elaborazione sia fermata, viene scritto il *data-set* ERRORI_INPUT. Nel *data-set* dovranno essere riscontrabili due righe per ciascun codice della variabile “Strato” coinvolto nell’incoerenza riscontrata, ovvero una riga in cui compare il codice di strato in corrispondenza al codice della variabile “Tipo di disegno” pari a “0” e una riga in cui compare il codice di strato in corrispondenza al codice della variabile “Tipo di disegno” pari ad “1”.

2.3 I data-set con le informazioni su stime ed errori campionari

In questo paragrafo si descrive la struttura ed il significato delle variabili contenute nei *data-set* SAS generate dal software. Nella trattazione si fa riferimento ai nomi originali delle variabili anziché alle etichette (*label*) ad esse assegnate. Affinché compaiano questi nomi, quando si visiona un *data-set*, l’utente deve aprirlo e successivamente selezionare l’opzione “*Column names*” dal menù “*Vien*”. Di seguito sono elencati i *data-set* di output e le loro principali caratteristiche.

Una premessa necessaria riguarda le informazioni memorizzate nei data-set STRATO, TOTALE, TOTDIS0 e TOTDIS1. Come verrà successivamente descritto nel *paragrafo* 2.5, per facilitare la lettura dei dati di output del software e agevolare l’utente nelle eventuali operazioni di ricerca, sono stati creati alcuni data-set *di lavoro* che contengono le stesse informazioni, ma che è possibile trattare più facilmente. Le differenze sono relative ai soli campi descritti nella successiva tabella 2.17 mentre il resto delle informazioni non varia.

❑ STRATO

Questo *data-set* contiene un primo insieme di informazioni relativo a ciascuna stima di interesse ed alla corrispondente variabilità campionaria con riferimento:

- a ciascuno strato di unità primarie, per ogni sottoclasse all'interno dello strato, e per ogni variabile di interesse; in particolare il *data-set* contiene un insieme di record ciascuno dei quali si riferisce ad una combinazione delle variabili indicate nella seguente tabella:

Tabella 2.1 – Variabili che definiscono i record del data-set STRATO

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO VARIABILE DEL DATA-SET	RAPPRESENTAZIONE SIMBOLICA
STRATI	Strato	Codice di strato, dopo l'eventuale aggregazione degli strati, se questi vengono collassati	$h=1,\dots,H$
SOTTOCLA	Variabili di sottoclasse	Variabile di sottoclasse (è pari a "0" se le informazioni contenute nel record non considerano la suddivisione in sottoclassi)	$s=1,\dots,S$
MODSCL	Modalità sottoclasse	Modalità della sottoclasse s ; nel caso in cui SOTTOCLA è pari a "0", MODSCL è anch'essa nulla	$m_s=1,\dots,M_s$
VARIABIL	Variabili di interesse	Variabile di interesse	$v=1,\dots,V$
MODALITA	Modalità variabili di interesse	Modalità della variabile di interesse, se la variabile è qualitativa; assume sempre valore 1 nel caso in cui la variabile è quantitativa	$m_v=1,\dots,M_v$

- La generica combinazione (h, s, m_s, v, m_v) contiene, quindi, le informazioni (cfr. tabella 2) riferite alla stima del totale relativo alla modalità m_v della variabile di interesse v con riferimento alla modalità m_s della sottoclasse s all'interno dello strato h di unità primarie.
- a ciascuno strato di unità primarie, senza la suddivisione in sottoclassi, e per ogni variabile di interesse; in particolare tali informazioni sono contenute nei record in cui le variabili SOTTOCLA e MODSCL sono poste pari a "0". La generica combinazione $(h, s=0,$

$m_s=0$ v, m_v) contiene, quindi, le informazioni (cfr. tabella 2.2) riferite alla stima del totale relativo alla modalità m_v della variabile di interesse v dello strato h di unità primarie.

Le informazioni che sono riportate nel *data-set* con riferimento a ciascuna stima di interesse ed alla corrispondente variabilità campionaria sono descritte nella tabella 2.2.

Tabella 2.2 – Informazioni del data-set STRATO riferite a ciascun record

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
STIMA	Stima	Stima del totale
VARFIN	Varianze stimatore finale	Stima della varianza dello stimatore adottato (cfr. appendice A.3)
VARDIR	Varianze stimatore diretto	Stima della varianza dello stimatore diretto
VARCLA*	Varianze stimatore cluster	Stima della varianza dello stimatore adottato, ottenuta in base ad un metodo che utilizza i coefficienti di riporto finali senza utilizzare, tuttavia, l'espressione linearizzata relativa allo stimatore
COMUNI	-	Variabile di <i>utility</i>
TOTALE2	-	Variabile di <i>utility</i>

*Il metodo di calcolo utilizzato è quello del software CLUSTERS (Verma, Scott e O' Muirchertaigh, 1980).

Un secondo insieme di informazioni contenute nel *data-set* si riferisce, invece, alle sottoclassi all'interno di ciascuno strato. In particolare, dette informazioni riguardano il numero totale stimato di unità della popolazione ed il numero totale di unità del campione. Poiché per ogni combinazione (h, s, m_s) riferita alla modalità m_s della generica sottoclasse s all'interno dello strato h il *data-set* presenta più record riferiti alle diverse variabili di interesse v ($v=1, \dots, V$; $m_v=1, \dots, M_v$), le informazioni riferite ad una data combinazione (h, s, m_s) sono ripetute in modo identico per tutti i record (h, v, m_v, s, m_s) identificati da detta combinazione. Le informazioni in oggetto sono descritte nella seguente tabella.

Tabella 2.3 – Informazioni del data-set STRATO riferite a ciascuna sottoclasse all'interno di ogni strato

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
POPCL	Stima del totale unità elementari	Numero stimato di unità elementari nella sottoclasse all'interno dello strato, ottenuto sommando i coefficienti finali delle unità appartenenti a tale sottoclasse
CAMPCL	Numero unità elementari	Numero totale di unità elementari campione appartenenti alla sottoclasse all'interno dello strato

Un terzo insieme di informazioni contenute nel *data-set* si riferisce, invece, ai soli strati. In particolare dette informazioni riguardano le numerosità campionarie al livello di strato, il tipo di disegno adottato nello strato ed il dominio di studio a cui appartiene lo strato. Poiché, con riferimento a ciascuno strato h , il *data-set* presenta più record riferiti alle diverse combinazioni (v, m_v, s, m_s) le informazioni riferite ad ogni strato sono ripetute in modo identico per tutti i record (h, v, m_v, s, m_s) riferiti a tale strato; le informazioni in oggetto sono descritte nella seguente tabella.

Tabella 2.4 – Informazioni del data-set STRATO riferite allo strato

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
ARNAR	Tipo disegno	Codice "Tipo di disegno adottato" nello strato
DOMI	Popolaz. pianif. utiliz. per lo stimatore	Codice della popolazione pianificata di riferimento utilizzata per la definizione dei totali noti in base ai quali è stato definito lo stimatore
DOMST	Dominio pianificato	Codice di dominio pianificato in cui si trova lo strato
OSSERVAZ	Numero osservazioni	Numero totale di unità elementari selezionate nello strato
UP	Numero di u.p.	Numero di unità primarie selezionate nello strato
UF	Numero di u.f.	Numero di unità finali selezionate nello strato
POP	Popolazione	Numero stimato di unità elementari ottenuto sommando i coefficienti finali
POPST	-	Numero stimato di unità elementari ottenuto sommando i coefficienti diretti

❑ **TOTALE**

Questo *data-set* è costruito aggregando a livello di dominio pianificato alcune informazioni contenute nel *data-set* STRATO. Il *data-set* presenta un primo insieme di statistiche relative a ciascuna stima di interesse con riferimento:

- a ciascun dominio pianificato di unità primarie, per ogni sottoclasse all'interno del dominio pianificato, e per ogni variabile di interesse; in particolare il *data-set* contiene un insieme di record ciascuno dei quali si riferisce ad una combinazione delle variabili indicate nella seguente tabella:

Tabella 2.5 – Variabili che definiscono i record del data-set TOTALE

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET	RAPPRESENTAZIONE SIMBOLICA
DOMST [*] (o DOMSTN)	Dominio pianificato	Codice di dominio pianificato	$d=1, \dots, D$
SOTTOCLA	Variabili di sottoclasse	Variabile di sottoclasse (è pari a "0" se le informazioni contenute nel record non considerano la suddivisione in sottoclassi)	$s=1, \dots, S$
MODSCL	Modalità sottoclasse	Modalità della sottoclasse s , nel caso in cui SOTTOCLA è pari a "0", MODSCL è anch'essa nulla	$M_s=1, \dots, M_s$
VARIABIL	Variabili di interesse	Variabile d'interesse	$v=1, \dots, V$
MODALITA	Modalità di variabili di interesse	Indica la modalità della variabile di interesse, se la variabile è qualitativa; assume sempre valore 1 nel caso in cui la variabile è quantitativa	$m_v=1, \dots, M_v$

**DOMST compare nel caso di stime per variabili qualitative; DOMSTN compare, invece, nel caso di stime per variabili quantitative*

La generica combinazione (d, s, m_s, v, m_v) contiene, quindi, le informazioni (cfr. tabella 2.6) riferite alla stima del totale relativo alla modalità m_v della variabile di interesse v con riferimento alla modalità m_s della sottoclasse s all'interno del dominio pianificato d :

- a ciascun dominio pianificato, senza la suddivisione in sottoclassi, e per ogni variabile di interesse; in particolare tali informazioni sono contenute nei record in cui le variabili SOTTOCLA e MODSCL sono poste pari a “0”. La generica combinazione ($d, s=0, m_s=0, v, m_v$) contiene, quindi, le informazioni (cfr. tabella 2.6) riferite alla stima del totale relativo alla modalità m_v della variabile di interesse v del dominio pianificato d di unità primarie.

Le informazioni che sono riportate nel *data-set* con riferimento a ciascuna stima di interesse ed alla corrispondente variabilità campionaria sono descritte nella seguente tabella:

Tabella 2.6 – Informazioni del data-set TOTALE riferite a ciascun record

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
DEFT	Deft	Effetto del disegno di campionamento
EFFSTIM	Effetto stimatore	Effetto stimatore
ERRAS	Errore standard	Errore assoluto (o Errore standard)
ERRCL*	Errore cluster	Errore assoluto approssimato, ottenuto in base alla varianza VARCLA
ERREL	Errore relativo	Errore relativo (o Coefficiente di variazione)
ERRELPC	Errore relativo %	Errore relativo percentuale
LIMINF	Limite inferiore I. C.	Limite inferiore dell'intervallo di confidenza con probabilità pari a 0,95
LIMSUP	Limite superiore I. C.	Limite superiore dell'intervallo di confidenza con probabilità pari a 0,95
STIMA	Stima	Stima del totale
RHO	Correlaz. Intraclasse	Coefficiente di correlazione intraclasse
SQM	Scarto q. medio	Stima della deviazione standard della variabile
VARFIN	Varianza stimatore finale	Stima della varianza dello stimatore adottato (cfr. appendice A.3)
VARDIR	Varianza stimatore diretto	Stima della varianza dello stimatore diretto
VARCLA*	Varianza cluster	Stima della varianza dello stimatore adottato, ottenuta in base ad un metodo che utilizza i coefficienti di riporto finali senza utilizzare, tuttavia, l'espressione linearizzata relativa allo stimatore
VARSRs	Varianza S.R.S	Varianza del campione casuale semplice di confronto (utilizzata per il calcolo del denominatore del Deft)
COMUNI	-	Variabile di <i>utility</i>
TOTALE2	-	Variabile di <i>utility</i>

*Il metodo di calcolo utilizzato è quello del software CLUSTERS (Verma, Scott e O' Muirchertaigh, 1980).

Un secondo insieme di informazioni contenute nel *data-set* si riferisce, invece, alle sottoclassi all'interno di ciascun dominio pianificato. In particolare, dette informazioni riguardano il numero totale stimato di unità della popolazione ed il numero totale di unità del campione. Poiché per ogni combinazione (d, s, m_s) riferita alla modalità m_s della generica sottoclasse s all'interno del dominio pianificato d , il *data-set* presenta più record riferiti alle diverse variabili di interesse v ($v=1, \dots, V$; $m_v=1, \dots, M_v$), le informazioni relative ad una data combinazione (d, s, m_s) sono ripetute in modo identico per tutti i record (d, v, m_v, s, m_s) identificati da detta combinazione. Le informazioni in oggetto sono descritte nella seguente tabella.

Tabella 2.7 – Informazioni del data-set TOTALE riferite a ciascuna sottoclasse all'interno di ogni dominio pianificato

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
POPCL	Stima del totale unità elementari	Numero stimato di unità elementari nella sottoclasse all'interno del dominio pianificato, ottenuto sommando i coefficienti finali delle unità appartenenti a tale sottoclasse
CAMPCL	Numero unità elementari	Numero totale di unità elementari nella sottoclasse

Un terzo insieme di informazioni contenute nel *data-set* si riferisce ai soli domini pianificati. In particolare dette informazioni riguardano le numerosità campionarie a livello di dominio. Poiché, con riferimento a ciascun dominio pianificato d , il *data-set* presenta più record riferiti alle diverse combinazioni (v, m_v, s, m_s) le informazioni riferite ad ogni dominio pianificato sono ripetute in modo identico per tutti i record (v, m_v, s, m_s, d) riferiti a tale dominio; le informazioni in oggetto sono descritte nella seguente tabella.

Tabella 2.8 – Informazioni del data-set TOTALE riferite al dominio pianificato

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
OSSERVAZ	Numero di osservazioni	Numero di unità elementari di campionamento
UP	Numero di u.p.	Numero di unità primarie di campionamento
UF	Numero di u.f.	Numero di unità finali di campionamento
POP	Popolazione	Numero stimato di unità elementari ottenuto sommando i coefficienti finali
B	Numero medio per u.p.	Numero medio di unità elementari per unità primaria

□ TOTDIS1 e TOTDIS0

Altri *data-set* generati dal software sono: TOTDIS1 ed TOTDIS0, aventi la stessa struttura del *data-set* TOTALE. Il *data-set* TOTDIS1 viene costruito secondo le medesime modalità adottate per TOTALE, ma utilizzando solo le informazioni relative agli strati identificati nel *data-set* in STRATO dalla variabile ARNAR pari a “1”. Per tali strati si adotta un disegno di campionamento ad uno stadio con probabilità di inclusione costante e senza reimmissione. Il *data-set* TOTDIS0 viene invece costruito utilizzando solo le informazioni relative agli strati in cui si adottano i restanti disegni di campionamento implementati dal software. Tali strati sono identificati nel *data-set* STRATO con la variabile ARNAR pari a “0”.

I precedenti *data-set* sono utili unicamente per lo studio di disegni campionari di tipo composito in cui le unità appartenenti a differenti strati possono essere selezionate in base a differenti disegni campionari. Ad esempio nelle indagini ISTAT sulle famiglie le unità appartenenti ai comuni auto rappresentativi vengono selezionate mediante un disegno ad uno stadio stratificato senza reimmissione delle unità, mentre le unità appartenenti ai comuni non auto rappresentativi sono estratte utilizzando un disegno a due stadi di selezione con stratificazione delle unità primarie. Per tale disegno di tipo composito l'utilizzo dei due *data-set* consente di scomporre l'effetto del disegno di campionamento ed altre importanti statistiche nelle due componenti dovute rispettivamente al disegno ad uno stadio e al disegno a due stadi.

Per il contenuto dei due *data-set* si rimanda alle tabelle 2.5, 2.6 e 2.7.

□ MODEL

Il *data-set* contiene per ciascun dominio pianificato i parametri del modello di regressione per la presentazione sintetica degli errori campionari utilizzato per la stima di frequenze (cfr. formula (A.5.23) dell'*appendice A.5*). Le informazioni del *data-set* sono contenute nella tabella 2.9.

Tabella 2.9 – Informazioni del data-set MODEL

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
DOMST (o DOMSTN)*	Codice di dominio pianificato. Identifica i record del <i>data-set</i>
A	Valore stimato del parametro α_1 del modello (A.5.23)
B	Valore stimato del parametro α_2 del modello (A.5.23)
R2	Indice di determinazione (R^2 %) del modello (A.5.23)

*DOMST compare nel caso di stime per variabili qualitative; DOMSTN compare, invece, nel caso di stime per variabili quantitative

❑ MODEL2

Il *data-set* contiene per ciascun dominio pianificato i parametri del modello di regressione per la presentazione sintetica degli errori campionari utilizzato per la stima di totali di variabili quantitative (cfr. formula (A.5.33) dell'appendice A.5). Le informazioni del *data-set* sono contenute nella tabella 2.10.

Tabella 2.10 – Informazioni del data-set MODEL2

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
DOMST (o DOMSTN)*	Codice di dominio di stima. Identifica i record del <i>data-set</i>
A	Valore stimato del parametro α_1 del modello (A.5.33)
B	Valore stimato del parametro α_2 del modello (A.5.33)
C	Valore stimato del parametro α_3 del modello (A.5.33)
R2	Indice di determinazione (R^2 %) del modello (A.5.33)

*DOMST compare nel caso di stime per variabili qualitative; DOMSTN compare, invece, nel caso di stime per variabili quantitative

❑ INTERP

Il *data-set* contiene le informazioni sugli errori campionari interpolati ottenuti in base al modello di regressione per la presentazione sintetica degli errori campionari utilizzato per la stima di frequenze. Il *data-set* presenta per ciascun dominio pianificato e per il totale della popolazione una serie di record ciascuno dei quali è riferito ad uno dei seguenti valori della variabile PERC: 0,1, 0,5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50. Ognuno dei valori della variabile PERC rappresenta un valore pre-

fissato di una stima di frequenze in termini percentuali; ad esempio il valore 20 indica una frequenza del 20%. Le variabili che identificano i record del *data-set* sono descritte nella tabella 2.11.

Tabella 2.11 – Variabili che definiscono i record del data-set INTERP

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
DOMST (o DOMSTN)*	Codice di dominio di stima. Il codice relativo al totale di popolazione è posto pari a TOTALE
PERC	Valori prefissati delle stime di frequenze percentuali: 0,1, 0,5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50

*DOMST compare nel caso di stime per variabili qualitative; DOMSTN compare, invece, nel caso di stime per variabili quantitative

Per il generico record, la variabile STIMA utilizzata nel modello (cfr. formula (A.5.23) dell'appendice A.5) come variabile esplicativa si ottiene moltiplicando il valore della variabile PERC per il valore della variabile MAXI che rappresenta il numero stimato di unità elementari ottenuto sommando i coefficienti finali. Le informazioni contenute nel *data-set* sono riportate nella seguente tabella.

Tabella 2.12 – Informazioni del data-set INTERP

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
MAXI	Numero stimato di unità elementari ottenuto sommando i coefficienti finali
STIMA	Valore prefissato della stima di frequenza, ottenuta come: $MAXI \times PERC$
R2	Indice di determinazione (R^2 %) del modello (A.5.23)
A	Valore stimato del parametro α_1 del modello (A.5.23)
B	Valore stimato del parametro α_2 del modello (A.5.23)
X1	Ottenuta come: logaritmo naturale di (STIMA/100)
Y1INT	Ottenuta come: $A + (B \times X1)$
ERR2INT	Ottenuta come: esponenziale di Y1INT
ERRINT	Ottenuta come: radice quadrata di ERR2INT
ERRINTP	Ottenuta come: $ERRINT \times 100$

❑ INTERP2

Il *data-set* contiene le informazioni sugli errori campionari interpolati ottenuti in base al modello di regressione per la presentazione sintetica degli

errori campionari utilizzato per la stima di totali di variabili quantitative (cfr. formula (A.5.33) dell'appendice A.5). Il *data-set* presenta per ciascun dominio pianificato e per il totale della popolazione una serie di record ciascuno dei quali è riferito ad uno dei seguenti valori della variabile PERC: 0,01; 0,02; 0,03; 0,04; 0,05; 0,1; 0,5; 1; 2; 3; 4; 5; 10; 15; 20; 25; 30; 35; 40; 45; 50. Le variabili che identificano i record del *data-set* sono descritte nella tabella 2.13.

Tabella 2.13 – Variabili che definiscono i record del data-set INTERP2

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
DOMST (o DOMSTN)*	Codice di dominio di stima. Il codice relativo al totale di popolazione è posto pari a TOTALE
PERC	Valori prefissati delle stime di frequenze percentuali: 0,01; 0,02; 0,03; 0,04; 0,05; 0,1; 0,5; 1; 2; 3; 4; 5; 10; 15; 20; 25; 30; 35; 40; 45; 50

*DOMST compare nel caso di stime per variabili qualitative; DOMSTN compare, invece, nel caso di stime per variabili quantitative

Moltiplicando ciascun valore della variabile PERC per la variabile MAXI si ottiene il corrispondente valore della variabile STIMA, che rappresenta la variabile esplicativa utilizzata per calcolare gli errori interpolati mediante il modello. Per ciascun dominio pianificato, la variabile MAXI è posta pari alla stima più elevata. Le informazioni contenute nel *data-set* sono riportate nella seguente tabella.

Tabella 2.14 – Informazioni del data-set INTERP2

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
MAXI	Stima più elevata tra quelle calcolate per le diverse variabili di interesse
STIMA	Valore prefissato della stima del totale, ottenuta come: $MAXI \times PERC$
R2	Indice di determinazione (R^2 %) del modello (A.5.33)
A	Valore stimato del parametro α_1 del modello (A.5.33)
B	Valore stimato del parametro α_2 del modello (A.5.33)
C	Valore stimato del parametro α_3 del modello (A.5.33)
X1	Uguale a STIMA
X2	Ottenuta come: quadrato di STIMA
Y1INT	Ottenuta come: $(A+B \times X1+C \times X2)/STIMA$
ERR2INT	Ottenuta come: quadrato di Y1INT
ERRINT	Uguale a: Y1INT
ERRINTP	Ottenuta come: $ERRINT \times 100$

2.4 I data-set con le informazioni sulla stratificazione e sul campione

□ TAB1

Le più importanti informazioni contenute in questo *data-set* riguardano il processo di *collassamento* degli strati (per approfondimento cfr. *paragrafo 1.1.2*). Per ogni record, che rappresenta uno strato originale del disegno, il software genera una serie di variabili descritte nella tabella 2.15.

Tabella 2.15 – Informazioni del data-set TAB1

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
AR	Tipo disegno	Codice "Tipo di disegno adottato" nello strato
CAMP_UF	Totale rek	Numero totale di unità elementari selezionate nello strato
CAMP_UP	Totale u.p.	Numero di unità primarie selezionate nello strato
CAMP_COD	Totale unità finale	Numero di unità finali selezionate nello strato
STRATO	Strato orig.	Codice originale dello strato del disegno, i codici di questa variabile sono progressivi ed univoci nell'ambito di ciascun valore della variabile DOMINIO
STRATON	Super strato	Codice superstrato in cui si trova lo strato originale; i codici di questa variabile sono progressivi ed univoci nell'ambito di ciascun valore della variabile DOMXX
FLAG	Tipo aggreg.	Indicatore di <i>collassamento</i> degli strati: pari a "0" per gli strati che non devono essere collassati; pari a "1" per gli strati collassati; pari a "2" per gli strati che devono essere collassati ma che non è stato possibile collassare
COEF	Stima per unità finale	Numero stimato di unità finali nello strato originale, ottenuto sommando i coefficienti diretti
COEF1	Stima per record	Numero stimato di unità elementari nello strato originale, ottenuto sommando i coefficienti diretti
DOMINIO	Popolaz. Pianif. Utiliz. per lo stimatore	Codice della popolazione pianificata utilizzata per definire lo stimatore
DOMXX	-	Codice generato dal concatenamento delle variabili DOMINIO e DOMSTIMA

Tabella 2.15 segue – Informazioni del data-set TAB1

DOMSTIMA	-	Codice del dominio pianificato
S_UP	-	Numero di unità primarie nel superstrato a cui lo strato originale è stato aggregato; nel caso in cui lo strato non è stato collassato coincide con il valore della variabile CAMP_UP
S_COEF	-	Numero stimato di unità finali nel superstrato a cui lo strato originale è stato aggregato, ottenuto sommando i coefficienti iniziali; nel caso in cui lo strato non è stato collassato coincide con il valore della variabile COEF
STRATFIN	-	Codice di superstrato in cui si trova lo strato originale; i codici di questa variabile sono progressivi ed univoci nell'ambito di ciascun valore della variabile DOMINIO
NCOM	-	Variabile di <i>utility</i>
CONT	-	Variabile di <i>utility</i>
SUPSTRA	-	Variabile di <i>utility</i>

❑ UNIC

Nel *data-set* sono riportati gli strati che la procedura non è stata in grado di collassare nonostante tale operazione fosse necessaria per stimare la varianza. I record, sono identificati dai codici degli strati originali non aggregati e presentano alcune informazioni illustrate nella tabella 2.16.

Tabella 2.16 – Informazioni del data-set UNIC

NOME VARIABILE	ETICHETTA VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
STRATO	Strato originale	Codice originale dello strato del disegno, i codici di questa variabile sono progressivi ed univoci nell'ambito di ciascun valore della variabile DOMINIO
DOMINIO	Pop. Pianif. Util. Per lo stimatore	Codice della popolazione pianificata di riferimento utilizzata per la definizione dei totali noti in base ai quali è stato definito lo stimatore
DOMSTIMA	Dominio pianificato	Codice del dominio pianificato
COEF	Stima pop. u. f. con pesi diretti	Numero stimato di unità finali nello strato originale, ottenuto sommando i coefficienti iniziali

2.5 I data-set con informazioni per elaborazioni successive e file di output

❑ Data-set

I data-set STRATO, TOTALE, TOTDIS0 e TOTDIS1 - analizzati nel *paragrafo 2.3* – sono creati dal software principalmente allo scopo di visualizzare e stampare le informazioni di output del software; ciò richiede la presenza di alcuni formati che, per essere gestiti in modo corretto, necessitano una competenza SAS non richiesta per alcun altro motivo. Dato che le informazioni memorizzate in tali data-set sono utili allo sviluppo di elaborazioni successive, per facilitare la lettura dei dati di output e agevolare eventuali operazioni di ricerca, sono stati creati alcuni data-set *di lavoro* che memorizzano le stesse informazioni dei data-set di cui sopra, ma sono più facili da utilizzare.

I data-set *di lavoro* sono WSTRATO, WTOTALE, WTOTDIS0 e WTOTDIS1. Le differenze rispetto ai data-set STRATO, TOTALE, TOTDIS0 e TOTDIS1 sono riscontrabili nelle informazioni riportate in tabella 2.17. Le variabili XVARIABIL e XSOTTOCLA sostituiscono le rispettive VARIABIL e SOTTOCLA e rappresentano le variabili di interesse e di sottoclasse ma sono definite entrambe da un numero progressivo. LABVAR e LABSOTTOCLA mettono in corrispondenza i progressivi con i nomi delle variabili originarie definite dall'utente.

Tabella 2.17 – Informazioni dei data set di lavoro

NOME VARIABILE	SIGNIFICATO DELLA VARIABILE DEL DATA-SET
XVARIABIL	Numero progressivo della variabile d'interesse
LABVAR	Nome della variabile di interesse
XSOTTOCLA	Numero progressivo di riferimento della variabile di sottoclasse (è pari a "0" se le informazioni contenute nel record non considerano la suddivisione in sottoclassi)
LABSOTTOCLA	Nome della variabile di sottoclasse (se xsottocla è pari a "0" allora labsottocla non contiene alcuna informazione)

SEZIONE III

**Un esempio di utilizzo della funzione
di Stima ed Errori campionari**

1. L'applicazione della funzione di Stime ed Errori campionari di Genesees V. 3.0

***Sintesi:** In questa sezione sono illustrati i passi descritti nelle Sezioni I e II precedenti con riferimento ad un data-set di esempio, analizzando inoltre i risultati ottenuti dall'applicazione della funzione di Stime ed Errori Campionari del software Genesees V. 3.0.*

L'utente, utilizzando il data-set esempio.sas7bdat memorizzato nella cartella c:\genesees\esempi di installazione, può ripetere le operazioni che seguono.

In questo capitolo si vuole mostrare un esempio di applicazione del software Genesees per calcolare la stima della varianza in un campione di unità estratte con differenti disegni di campionamento.

Per far ciò è necessario specificare opportunamente l'input; in particolare, per ciascuno dei record del data-set SAS, è necessario definire le variabili di input connesse con il disegno ("Tipo di disegno adottato", "Unità primaria", "Unità finale"- cfr. *paragrafo 1.1.1, Sezione II*), in modo da distinguere se le unità campionarie sono state estratte secondo un disegno ad uno stadio piuttosto che secondo un disegno a due o più stadi.

Per agevolare l'utente vengono specificati i riferimenti alle due precedenti sezioni: la *Sezione I* è utile per l'utilizzo delle interfacce; la *Sezione II* è utile per la costruzione dell'input. Sono dunque descritti i passi da seguire per effettuare quanto riportato nelle due precedenti sezioni, con riferimento ad un data-set di esempio e analizzando infine i risultati ottenuti dall'applicazione del software.

L'utente, utilizzando il data-set esempio.sas7bdat memorizzato nella cartella c:\genesees\esempi di installazione, può ripetere le operazioni che seguono.

Per analizzare la costruzione del data-set tramite un esempio di applicazione del software a un data-set, si legga il paragrafo 1.1 in questa sezione.

Per seguire come utilizzare la funzione di Stime ed Errori del software Genesees V. 3.0 e per l'analisi degli output, si legga il paragrafo 1.2.

1.1. La costruzione del data-set di input

Come descritto nella *Sezione II*, il data-set deve essere costruito in base al tipo di stimatore, al disegno campionario adottato e al livello di stima considerato.

Il data-set esempio.sas7bdat (cfr. *figura 1.1*) ha una struttura analoga a quella del data-set di input utilizzato dall'ISTAT per calcolare le stime e gli errori campionari dell'indagine Forze Lavoro. Poiché tale data-set serve unicamente ad illustrare l'utilizzo della procedura in un caso concreto, in cui il disegno campionario alla base dell'indagine è di tipo *composto*, i singoli valori riportati nel data-set esempio.sas7bdat sono del tutto fittizi.

Affinché risultino chiari i successivi passi, è utile riassumere alcune informazioni sull'indagine sulle forze di lavoro condotta dall'ISTAT.

La detta indagine è una rilevazione trimestrale sulle famiglie. La popolazione d'interesse è costituita da tutti gli individui residenti in Italia, al netto dei membri permanenti delle convivenze. L'unità di rilevazione è la famiglia anagrafica.

Il disegno di campionamento è di tipo composto e prevede la stratificazione dei comuni che costituiscono le unità primarie di campionamento.

La stratificazione dei comuni viene effettuata all'interno di ogni provincia in base alla dimensione demografica: in ciascuna provincia i comuni sono suddivisi in due sottoinsiemi, i comuni di maggiore dimensione demografica costituiscono uno strato a sé stante e sono definiti Auto Rappresentativi, i rimanenti comuni sono definiti Non Auto Rappresentativi e sono ulteriormente stratificati in modo da costituire strati di uguale ampiezza demografica.

I primi vengono tutti inclusi nel campione, mentre da ognuno degli strati Non Auto Rappresentativi vengono selezionati senza reimmissione due comuni con probabilità proporzionale alla dimensione demografica. Il disegno di campionamento è quindi di tipo *composto*, prevedendo uno stadio di selezione negli strati Auto Rappresentativi, dai quali sono selezionate direttamente le *famiglie* (unità primarie), e due stadi di selezione negli strati Non Auto Rappresentativi, dai quali sono selezionati due *comuni* (unità primarie) e successivamente le *famiglie* (unità finali). *Tutti i componenti* riportati nel foglio di famiglia vengono sottoposti a rilevazione.

1.1.1 Il data-set di esempio

Figura 1.1: Il data-set esempio.sas7bdat

COD_FA	PESO_DIS	PESO_FIN	PESO_INIZ	PROVIN	REGIO	SE	STRATO	TIPO_D	UN_PRIM	X1	Y1	▲
1	1	34.3	37	PROV1	REG1	2	str01	1	1	0	C	
2	1	31.1	37	PROV1	REG1	1	str01	1	2	0	C	
2	1	31.1	37	PROV1	REG1	2	str01	1	2	0	C	
3	1	30.2	37	PROV1	REG1	1	str01	1	3	0	C	
3	1	30.2	37	PROV1	REG1	2	str01	1	3	0	C	
4	1	28.8	37	PROV1	REG1	1	str01	1	4	0	C	
4	1	28.8	37	PROV1	REG1	2	str01	1	4	0	C	
5	1	40.5	37	PROV1	REG1	1	str01	1	5	0	C	

Le variabili del data-set SAS esempio.sas7bdat, relative a dati individui, sono state costruite secondo i formati conformi a quanto definito nel *paragrafo 1.1 della Sezione II*:

UN_PRIM	<i>type=number</i>	<i>length=8</i>
COD_FAM	<i>type=number</i>	<i>length=8</i>
PESO_INIZ	<i>type=number</i>	<i>length=8</i>
PESO_FIN	<i>type=number</i>	<i>length=8</i>
REGIONE	<i>type=text</i>	<i>length=4</i>
PROVINCIA	<i>type=text</i>	<i>length=5</i>
SEX	<i>type=text</i>	<i>length=1</i>
TIPO_DISE	<i>type=text</i>	<i>length=1</i>
PESO_DIST	<i>type=number</i>	<i>length=8</i>
STRATO	<i>type=text</i>	<i>length=5</i>
Y1-Y6	<i>type=number</i>	<i>length=8</i>
X1-X50	<i>type=number</i>	<i>length=8</i>

In particolare, le variabili di interesse Y1-Y6 rappresentano:

- Y1 l'appartenenza alle forze di lavoro;
- Y2 l'appartenenza all'insieme degli occupati;
- Y3 l'appartenenza all'insieme delle persone in cerca di occupazione;
- Y4 l'appartenenza all'insieme delle persone in cerca di prima occupazione;
- Y5 l'appartenenza all'insieme dei disoccupati;
- Y6 l'appartenenza all'insieme delle altre persone in cerca di occupazione.

Le variabili sono di tipo *qualitativo* ed assumono valore 1 o 0. La variabile Y1, ad esempio, assume valore pari a 1, se l'individuo appartiene alle forze di lavoro, pari a 0, altrimenti.

Il significato delle altre variabili è chiarito nella tabella 1.1.

Tabella 1.1 Le variabili di input

Variabile di input	Variabile nel data-set SAS di input
Tipo di disegno adottato	TIPO_DISE
Unità 1	UN_PRIM
Unità 2	COD_FAM
Strato	STRATO
Peso diretto	PESO_INIZ
Peso finale	PESO_FIN
Variabili ausiliarie	X1, ..., Xj, ..., X50
Popolazione pianificata utilizzata per lo stimatore	REGIONE
Peso distanza	PESO_DIST
Variabili di interesse	Y1, ..., Y6
Variabile di sottoclasse	SEX
Dominio pianificato	PROV

1.1.2 La costruzione delle variabili di input

La stima della varianza campionaria deve essere calcolata tenendo conto dei differenti disegni di campionamento utilizzati per l'estrazione delle unità.

Come descritto nel paragrafo 1.1, la parte autorappresentativa del campione presuppone un **campionamento stratificato di grappoli di unità elementari senza reimmissione e con probabilità di inclusione costante**.

Il data-set di input è stato costruito seguendo le indicazioni riportate nel paragrafo 1.3.2.4 della *Sezione II*, ne segue che:

- a) la variabile TIPO_DISE (corrispondente alla variabile TIPO_DIS del paragrafo 1.3.2.4, Sezione II) deve assumere valore pari ad “1” su tutti i record;
- b) la variabile STRATO deve assumere tanti valori distinti quanti sono gli strati;
- c) la variabile STRATO deve assumere un valore uguale per tutti i record appartenenti a grappoli contenuti nello stesso strato; record appartenenti a grappoli contenuti in strati diversi presentano valori diversi della variabile STRATO;
- d) per garantire che effettivamente i valori risultino distinti conviene sempre assegnare alla variabile STRATO il valore della numerazione progressiva degli strati dell’intero data-set; nel data-set esempio.sas7bdat la variabile assume 17 valori distinti, identificati con i codici alfanumerici “str01”, “str02”,... , “str17”;
- e) all’interno di uno strato la variabile COD_FAM (corrispondente alla variabile UNITA_2 del paragrafo 1.3.2.4, Sezione II) deve assumere tanti valori distinti pari al numero di grappoli che appartengono allo strato stesso. A ciascun valore della variabile COD_FAM è associato un grappolo di unità elementari (nel data-set esempio.sas7bdat la famiglia), quindi, all’interno di uno strato i record appartenenti allo stesso grappolo (gli individui) presentano lo stesso valore della variabile COD_FAM. Record appartenenti a grappoli diversi contenuti in uno stesso strato devono avere un diverso valore della variabile COD_FAM;
- f) la variabile UN_PRIM (corrispondente alla variabile UNITA_1 del paragrafo 1.3.2.4, Sezione II) assume i medesimi valori della variabile COD_FAM.

Il data-set di input relativo alla parte non autorappresentativa, nella quale il **campionamento è a due stadi**, è stato costruito procedendo come descritto nel paragrafo 1.3.2.7, *Sezione II* e pertanto:

- g) la variabile TIPO_DISE deve assumere valore pari a “0” su tutti i record;

- h) la variabile STRATO assume tanti valori distinti quanti sono gli strati;
- i) la variabile STRATO deve assumere un valore uguale per tutti i record appartenenti a unità primarie - identificate dai codici della variabile UN_PRIM, (si veda il punto successivo) contenute nello stesso strato; record appartenenti a unità primarie contenute in strati diversi presentano valori diversi della variabile STRATO;
- j) per la determinazione della variabile STRATO si suggerisce di procedere come descritto nel punto (d);
- k) all'interno di uno strato la variabile UN_PRIM assume tanti valori distinti pari al numero di unità primarie che appartengono allo strato stesso. A ciascun valore della variabile UN_PRIM è associata una unità primaria (nel caso in esame, un comune) e quindi i record appartenenti allo stesso comune NAR presentano lo stesso valore della variabile UN_PRIM. Record appartenenti a comuni diversi devono avere un diverso valore della variabile UN_PRIM;
- l) è necessario controllare che nella parte non auto rappresentativa non ci siano valori della variabili UN_PRIM uguali a quelli relativi a record della parte autorappresentativa;
- m) all'interno di una unità primaria la variabile COD_FAM (corrispondente alla variabile UNITA_2 del *paragrafo 1.3.2.7, Sezione II*) assume tanti valori distinti pari al numero di unità secondarie (record) che appartengono alla stessa unità primaria. A ciascun valore della variabile COD_FAM è associata una unità secondaria.

Lo stimatore utilizzato nell'indagine delle forze di lavoro è uno stimatore di calibrazione, in cui le informazioni ausiliarie utilizzate sono i totali di popolazione per sesso e per 14 classi di età, relativi alla regione, e i totali per sesso, relativi alla provincia.

La specificazione delle variabili di input sulla base dello stimatore adottato è stata descritta nel *paragrafo 1.3.1*. Nel caso in esame è necessario specificare il livello del modello (*paragrafo 1.3.1.3, Sezione II*) a *livello di unità elementare*, in quanto sia le variabili di interesse sia quelle ausiliarie sono relative a ciascun elemento della popolazione.

La definizione delle variabili del data-set relative al tipo di modello adottato avviene, nel caso in cui si utilizza lo stimatore di calibrazione (*paragrafo 1.3.1.4, Sezione II*), secondo lo schema presentato in figura 1.2:

Figura 1.2 : la costruzione delle variabili del data-set SAS

Regione	un_ prim	cod- fam	X1	X2	X3..X27	X28	X29	X30	X31..X49	X50	peso dist
Reg1	1	0	0.....0	0	1	0	0.....0	0	1
...
Reg1	0	1	0.....0	0	0	1	0.....0	0	1
.....
Reg2	0	0	0...1...0	0	0	0	0...1.....0	0	1
.....
.....
.....

Più precisamente, le variabili sono definite come segue:

- la variabile “Popolazioni pianificate utilizzate per lo stimatore” è identificata dalla regione di appartenenza dell’individuo;
- le variabili ausiliarie X_j sono distinte in due gruppi: X_1 - X_{28} e X_{29} - X_{50} : nel primo le variabili, identificate da un numero progressivo da 1 a 28, sono indicatrici delle combinazioni di modalità delle variabili “sesso” (due modalità) e “classe di età” (quattordici modalità). Pertanto, per ogni record tutte le variabili X_j sono nulle tranne per quella associata alle modalità di sesso e classe di età, che identificano la sottopopolazione a cui l’unità elementare considerata appartiene. Tale variabile assume valore pari a 1. A titolo esemplificativo, la sottopopolazione in cui si trova il primo record della figura 1.2 corrisponde a quella definita da $X_1=1$, ovvero la popolazione dei maschi appartenenti alla prima classe di età. Il secondo gruppo di variabili X_{29} - X_{50} rappresenta, invece, le variabili indicatrici dei caratteri sesso e provincia (entro la regione). Pertanto, essendo il numero massimo di province all’interno di una regione pari a 11, X_{29} sarà pari a 1 se l’individuo è di sesso maschile e se appartiene alla prima provincia della regione, 0 altrimenti; X_{30} è pari a 1 se l’individuo è di sesso femminile e se appartiene alla prima provincia della regione, 0 altrimenti; X_{31} è pari a 1 se l’individuo è di sesso maschile e se appartiene alla

seconda provincia della regione, 0 altrimenti; X32 è pari a 1 se l'individuo è di sesso femminile e appartiene alla seconda provincia della regione, 0 altrimenti e così via (le regioni che hanno meno di 11 province avranno variabili X con valori tutti nulli a partire da un certo indice in poi);

- c) alla variabile PESO_DIST (il peso CK nella descrizione dello stimatore di calibrazione), corrispondente al peso assegnato a ciascuna unità, viene assegnato valore pari a 1;
- d) la variabile PESO_FIN è il peso finale da attribuire alle unità con riferimento al tipo di stimatore e ai vincoli adottati. La sua determinazione può essere effettuata utilizzando la funzione di Riponderazione del software, al cui manuale si rimanda per una descrizione particolareggiata (Pagliuca, 2004a).

Nel caso qui esaminato, l'obiettivo è quello di valutare gli errori delle stime a livello provinciale e per sesso e classe d'età in ciascuna provincia.

Poiché il disegno di campionamento prevede che la stratificazione dei comuni sia effettuata all'interno di ogni provincia, è possibile selezionare la variabile PROV (provincia) come variabile "Dominio Pianificato". Poiché, invece, la variabile SEX (sesso) non costituisce un dominio pianificato, essa deve essere selezionata tra le variabili di sottoclasse.

Per determinare la stima degli errori a livello regionale è possibile selezionare anche la variabile regione tra le variabili di sottoclasse; oppure si può implementare nuovamente il software imponendo che la regione sia il "Dominio Pianificato"; o alternativamente si può ottenere come somma degli errori a livello provinciale per ogni singola regione.

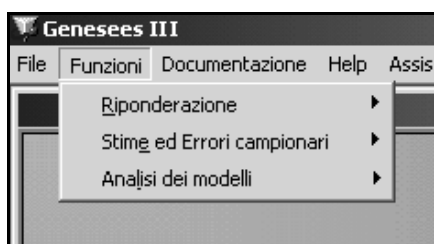
Gli errori relativi al totale nazionale sono automaticamente determinati dal software, come somma degli errori su tutti i domini pianificati.

1.2 L'uso del software e la presentazione dell'output

1.2.1 L'uso delle schermate utilizzando il data-set di esempio

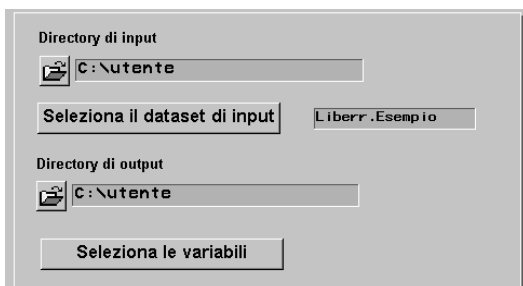
Come illustrato nel *capitolo 5, Sezione I*, all'avvio della procedura la schermata principale appare come in figura 1.3. Per la valutazione degli errori campionari l'utente deve selezionare la voce "Stime ed Errori campionari" del menu "Funzioni".

Figura 1.3 - La selezione della voce "Stime ed Errori campionari"



Nella maschera successiva (cfr. figura 1.4) si richiede all'utente di specificare sia il data-set contenente le informazioni necessarie per l'elaborazione sia la cartella di output, eventualmente coincidente con la cartella di input; nel nostro esempio **c:\utente** è la cartella in cui è contenuto il data-set esempio.sas7bdat e in cui viene memorizzato l'output.

Figura 1.4 - La selezione del data-set



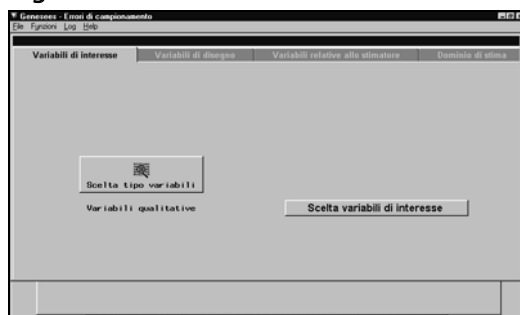
Utilizzare il bottone "Seleziona le variabili" per attivare la maschera successiva, composta di quattro schede distinte relative a quattro differenti

tipologie di variabili (figure 1.5, 1.7, 1.8 e 1.9). L'utente che voglia maggiori informazioni sulla selezione dei dati di input può ripercorrere quanto descritto nel *capitolo 5, Sezione I*, seguendo le illustrazioni lì presentate riferite allo stesso data-set di questa applicazione.

Variabili di interesse

La scheda relativa alla selezione delle “Variabili di interesse” è formata da due bottoni (cfr. figura 1.5).

Figura 1.5 – Maschera di selezione delle variabili di input - Variabile di interesse



Il primo bottone deve essere utilizzato per specificare che le variabili di interesse del data-set esempio.sas7bdat sono di tipo qualitativo. Il secondo bottone avvia una ulteriore maschera (cfr. figura 1.6) che permette di selezionare le variabili. Nel nostro esempio vengono scelte le variabili del data-set: Y1 Y2 Y3 Y4 Y5 Y6.

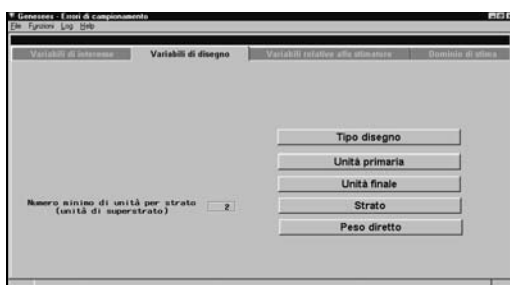
Figura 1.6 - Maschera di selezione delle variabili d'interesse



Variabili di disegno

Per indicare le variabili di disegno si deve attivare la seconda scheda (cfr. figura 1.7).

Figura 1.7 – Maschera di selezione delle variabili di input - Variabili di disegno



Ognuno dei cinque bottoni di questa scheda attiva una maschera che consente di specificare una sola variabile. Il campo editabile, che presenta un valore di default pari a due, permette di variare il numero minimo delle unità che si vuole aggregare in un eventuale processo di *collassamento*.

Per il significato delle variabili che è possibile selezionare si può leggere il *paragrafo 1.1.1. della Sezione II*; per ciò che concerne il *collassamento* si può leggere il *paragrafo 1.2.1 della Sezione II*.

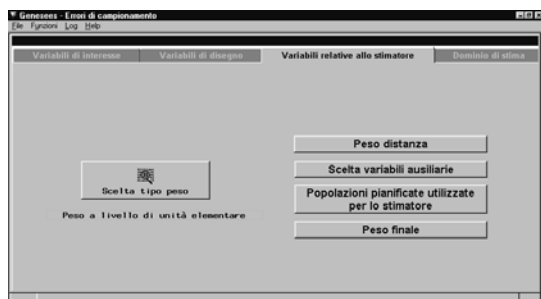
Nel nostro esempio vengono scelte le seguenti variabili del data-set:

- *TIPO_DISE* tramite il bottone “Tipo disegno”;
- *UN_PRIM* tramite il bottone “Unità primaria”;
- *COD_FAM* tramite il bottone “Unità finale”;
- *STRATO* tramite il bottone “Strato”;
- *PESO_INIZ* tramite il bottone “Peso diretto”.

Variabili relative allo stimatore

La terza scheda serve a selezionare le variabili relative allo stimatore (cfr. *figura 1.8*).

Figura 1.8 – Maschera di selezione delle variabili di input - Variabili relative allo stimatore



Il bottone “Scelta tipo peso” consente di specificare se i pesi del data-set sono *a livello di cluster* ovvero *a livello di unità elementare* (paragrafo 1.3.1.3, Sezione II); nel caso del data-set esempio.sas7bdat si dovrà utilizzare il bottone in modo che compaia la dicitura “Peso a livello di unità elementare”.

Ciascuno degli altri bottoni attiva una maschera per la selezione delle rispettive variabili; si ricorda che è possibile specificare una sola variabile, fatta eccezione per la maschera relativa alla selezione delle variabili ausiliarie.

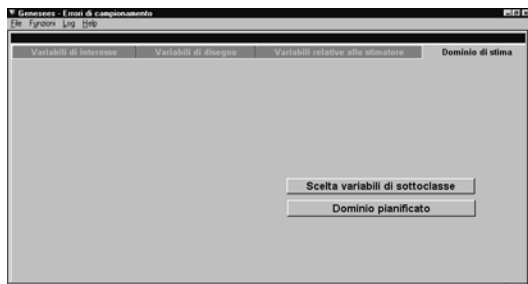
Nella applicazione, sono state scelte le seguenti variabili del data-set:

- *PESO_DIST* tramite il bottone “Scelta tipo peso”
- *X1-X50* tramite il bottone “Scelta variabili ausiliarie”
- *REGIONE* tramite il bottone “Popolazioni pianificate utilizzate per lo stimatore”
- *PESO_FIN* tramite il bottone “Peso finale”

Dominio di stima

La quarta scheda (cfr. figura 1.9) concerne la selezione delle variabili relative al dominio di stima e presenta i due bottoni da utilizzare per scegliere le “variabili di sottoclasse” (è consentito scegliere più variabili) o per scegliere la variabile corrispondente al “dominio pianificato” (si sceglierà una unica variabile) (cfr. paragrafo 1.1.1, Sezione II).

Figura 1.9 – Maschera di selezione delle variabili di input - Variabili relative al dominio di stima

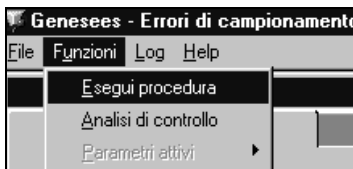


Nell'esempio considerato, sono state scelte le seguenti variabili del data-set:

- *SEX* tramite il bottone “Scelta variabili di sottoclasse”
- *PROVINCIA* tramite il bottone “Dominio pianificato”

Selezionate le variabili e i parametri di input, si può procedere con l'esecuzione della procedura per calcolare le stime e gli errori di campionamento, scegliendo la voce “Esegui procedura” del menu “Funzioni”:

Figura 1.10: La voce “Esegui procedura”



Si osservi che l'utente ha la possibilità di avviare la voce “Analisi controllo” ed effettuare la stampa di controllo a video dei dati di input (cfr. *paragrafo 5.2, Sezione I*) prima di eseguire la procedura.

Qualora non tutte le variabili obbligatorie siano state inserite, la procedura invia un messaggio di errore, altrimenti procede nella elaborazione.

Nella prima fase dell'esecuzione, gli strati costituiti da una sola unità campionaria sono sottoposti ad un processo automatico di *collassamento*. I *superstrati* prodotti hanno un numero di unità primarie il cui minimo è dato dal parametro di *collassamento* specificato dall'utente; nel caso in esame tale parametro è stato posto pari a due (cfr. *figura 1.7*).

A titolo esemplificativo, nel data-set sono stati inseriti due strati (str02 e str03) con una unica unità primaria ma che risultano non aggregabili.

Infatti, secondo quanto indicato nel *paragrafo 1.1.2 della Sezione II*, per formare i superstrati è necessario lasciare inalterati gli strati per i quali è possibile calcolare la varianza e, per quelli che invece presentano una unica unità primaria, devono verificarsi le seguenti condizioni :

- a) si devono aggregare strati simili e gli strati aggregati devono essere formati da unità che appartengono alla stessa popolazione pianificata utilizzata per lo stimatore (con riferimento all'esempio, la Regione);
- b) per rispettare il livello di stima finale desiderato, gli strati aggregati devono essere formati da unità che appartengono anche allo stesso dominio di stima pianificato (con riferimento all'esempio, la Provincia);
- c) per evitare di aggregare unità estratte secondo disegni diversi, gli strati aggregati devono essere formati da unità che presentano lo stesso valore della variabile "Tipo di disegno".

Come si può osservare analizzando i dati del data-set *esempio.sas7bdat*, lo strato str02 è formato da un unico comune Non Auto Rappresentativo ("Tipo di disegno"='0' per ogni record); il software dovrebbe aggregare automaticamente tale strato con un altro appartenente alla stessa popolazione pianificata utilizzata per lo stimatore REG1. Nel data-set in esame, con riferimento alla popolazione REG1, è possibile trovare oltre a str02, lo strato str01 che è formato da un unico comune Autorappresentativo ("Tipo di disegno"='1' per ogni strato), le cui unità sono dunque estratte con un diverso disegno. Per il precedente punto 3, l'aggregazione non è possibile.

Anche lo strato str07 presenta un problema analogo; con riferimento a REG2 e PROV3 (si vedano i punti 1 e 2 di cui sopra) gli strati str08 e str09 hanno lo stesso tipo di disegno, ma presentano un numero di unità primarie superiore ad 1. Pertanto non rientrano tra gli strati sottoposti al processo di aggregazione.

Il software mostra gli strati non aggregabili con una finestra (cfr. *figura 1.11*).

Figura 1.11: Finestra che appare dal processo di aggregazione in Superstrati

Strati con una sola unità primaria				
	strato originale	pop.pianif.util. per stimato	dominio pianificat	peso diretto ▲
1	str02	REG1	PROV1	555
2	str07	REG2	PROV3	7760

Il software invia un messaggio di avviso per segnalare gli strati che non sono stati aggregati e permette all'utente di proseguire. E' da tenere presente che nei domini pianificati che includono gli strati str02 e str07 la varianza risulterà sottostimata.

Dopo questo controllo il software procede con l'elaborazione richiesta.

Al termine dell'esecuzione appare il messaggio "l'elaborazione è terminata".

Nella seguente tabella 1.2 vengono riassunte le scelte effettuate per l'applicazione, così come è possibile visualizzare con la voce "Parametri attivi" (cfr. *paragrafo 5.2.3, Sezione I*):

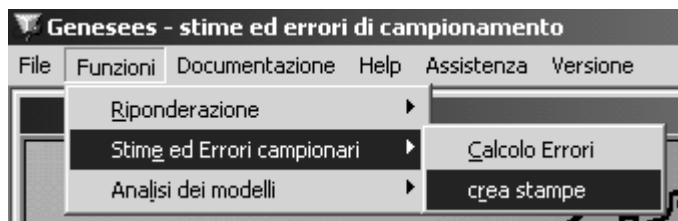
Tabella 1.2: La selezione delle variabili dal data-set esempio.sas7bdat

Variabili di interesse	Y1 Y2 Y3 Y4 Y5 Y6
Peso distanza	PESO_DIST
Variabili ausiliarie	X1 X2X50
Dominio pianificato	PROVINCIA
Peso finale	PESO_FIN
Sottoclassi	SEX
Tipo di disegno	TIPO_DISE
Unità primaria	UN_PRIM
Unità finale	COD_FAM
Strato	STRATO
Peso diretto	PESO_INIZ
Popolaz.pianif.util. per stimatore	REGIONE
Nr. unità superstrato	2
Variabili: 1 quant., 2 qual.	2
Stima: 1 unità elem., 2 cluster	1

1.2.2 Le stampe che si ottengono utilizzando il data-set di esempio

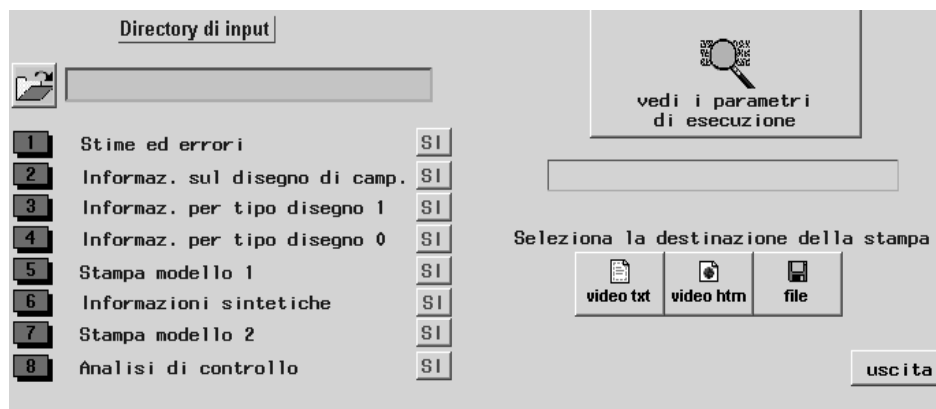
Per ottenere le stampe è necessario riposizionarsi sulla maschera della schermata principale (cfr. *figura 1.3*) e selezionare la voce “Crea stampe” (cfr. *figura 1.12*):

Figura 1.12 - La selezione della voce “Crea stampe”



Nel *paragrafo 5.3* della *Sezione I* si è descritta la voce “Crea stampe” specificando che l’utente può scegliere di visualizzare le stampe a video in formato ASCII (bottone “video.txt”), in formato HTML (bottone “video.htm”) o produrre dei file esterni per memorizzare le stampe (bottone “file”) (cfr. *figura 1.13*).

Figura 1.13 - Maschera di selezione delle stampe



I file ASCII - scritti solo se l'utente ha selezionato la stampa corrispondente - sono i seguenti: stampa1.txt, stampa2.txt, stampa3.txt, stampa4.txt, stampa5.txt, stampa6.txt, stampa7.txt, stampa8.txt.

Sempre nello stesso *paragrafo 5.3* della *Sezione I* sono state mostrate le stampe a video.

A titolo esemplificativo, vengono di seguito riportate le stampe prodotte dal software sulla base della elaborazione del data set esempio.sas7bdat.

Ricordiamo che il data set è del tutto fittizio e pertanto il risultato prodotto potrebbe risultare privo di significato statistico.

Per migliorare la stampa, i file in formato ASCII sono stati importati in Microsoft Word e il testo è stato convertirlo in SAS Monospace, punti 8.

Stampa 1

Dalla stampa 1 è possibile ricavare alcune informazioni di base circa il valore delle stime e le variabilità di queste.

(da stampa1.txt)

1 - Stime ed errori di campionamento per dominio di stima variabili qualitative								
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0								
variabili	modalità	errore			limite		limite	
di interesse	variabili interesse	stima	standard	relativo %	inf. I.C.	sup. I.C.		
Y1	0	318150.70	3617.56	1.14	311060.3	325241.1		
Y1	1	216604.10	3617.56	1.67	209513.7	223694.5		
Y2	0	336971.00	3764.99	1.12	329591.6	344350.4		
Y2	1	197783.80	3764.99	1.90	190404.4	205163.2		
Y3	0	472925.40	3835.30	0.81	465408.2	480442.6		
Y3	1	61829.40	3835.30	6.20	54312.20	69346.60		
Y4	0	398800.40	5026.44	1.26	388948.6	408652.2		
Y4	1	135954.40	5026.44	3.70	126102.6	145806.2		
Y5	0	521332.00	2187.62	0.42	517044.3	525619.7		
Y5	1	13422.80	2187.62	16.30	9135.06	17710.54		
Y6	0	468722.30	4655.55	0.99	459597.4	477847.2		
Y6	1	66032.50	4655.55	7.05	56907.63	75157.37		
dominio pianificato=TOTALE sottoclasse=SEX modalità di sottoclasse=1								
variabili	modalità	errore			limite		limite	
di interesse	variabili interesse	stima	standard	relativo %	inf. I.C.	sup. I.C.		
Y1	0	129918.00	2648.13	2.04	124727.7	135108.3		
Y1	1	134115.90	2648.13	1.97	128925.6	139306.2		
Y2	0	137352.50	2349.40	1.71	132747.7	141957.3		
Y2	1	126681.40	2349.40	1.85	122076.6	131286.2		
Y3	0	223517.30	3561.87	1.59	216536.0	230498.6		
Y3	1	40516.60	3561.87	8.79	33535.34	47497.86		
Y4	0	177869.10	3175.14	1.79	171645.8	184092.4		
Y4	1	86164.80	3175.14	3.68	79941.53	92388.07		
Y5	0	255622.90	1704.12	0.67	252282.8	258963.0		
Y5	1	8411.00	1704.12	20.26	5070.92	11751.08		
Y6	0	215781.30	2979.04	1.38	209942.4	221620.2		
Y6	1	48252.60	2979.04	6.17	42413.69	54091.51		

(segue da stampa1.txt)

dominio pianificato=TOTALE sottoclasse=SEX modalità di sottoclasse=2									
variabili di interesse	variabili di interesse	modalità di interesse	stima	errore standard	errore relativo %	limite inf. I.C.	limite sup. I.C.		
Y1	0		188232.70	3226.29	1.71	181909.2	194556.2		
Y1	1		82488.20	3226.29	3.91	76164.67	88811.73		
Y2	0		199618.50	3536.89	1.77	192686.2	206550.8		
Y2	1		71102.40	3536.89	4.97	64170.10	78034.70		
Y3	0		249408.10	1967.12	0.79	245552.5	253263.7		
Y3	1		21312.80	1967.12	9.23	17457.24	25168.36		
Y4	0		220931.30	2803.76	1.27	215435.9	226426.7		
Y4	1		49789.60	2803.76	5.63	44294.22	55284.98		
Y5	0		265709.10	909.04	0.34	263927.4	267490.8		
Y5	1		5011.80	909.04	18.14	3230.08	6793.52		
Y6	0		252941.00	2657.54	1.05	247732.2	258149.8		
Y6	1		17779.90	2657.54	14.95	12571.11	22988.69		

dominio pianificato=PROV1 sottoclasse=0 modalità di sottoclasse=0									
variabili di interesse	variabili di interesse	modalità di interesse	stima	errore standard	errore relativo %	limite inf. I.C.	limite sup. I.C.		
Y1	0		21477.40	359.46	1.67	20772.85	22181.95		
Y1	1		15180.60	359.46	2.37	14476.05	15885.15		
Y2	0		22678.10	378.22	1.67	21936.80	23419.40		
Y2	1		13979.90	378.22	2.71	13238.60	14721.20		
Y3	0		32997.70	367.30	1.11	32277.79	33717.61		
Y3	1		3660.30	367.30	10.03	2940.39	4380.21		
Y4	0		26338.40	434.93	1.65	25485.94	27190.86		
Y4	1		10319.60	434.93	4.21	9467.14	11172.06		
Y5	0		36260.50	100.60	0.28	36063.32	36457.68		
Y5	1		397.50	100.60	25.31	200.32	594.68		
Y6	0		33937.50	280.33	0.83	33388.06	34486.94		
Y6	1		2720.50	280.33	10.30	2171.06	3269.94		

dominio pianificato=PROV1 sottoclasse=SEX modalità di sottoclasse=1									
variabili di interesse	variabili di interesse	modalità di interesse	stima	errore standard	errore relativo %	limite inf. I.C.	limite sup. I.C.		
Y1	0		9091.40	233.99	2.57	8632.78	9550.02		
Y1	1		8752.60	233.99	2.67	8293.98	9211.22		
Y2	0		9795.20	242.27	2.47	9320.36	10270.04		
Y2	1		8048.80	242.27	3.01	7573.96	8523.64		
Y3	0		15389.90	252.60	1.64	14894.80	15885.00		
Y3	1		2454.10	252.60	10.29	1959.00	2949.20		
Y4	0		12249.30	302.96	2.47	11655.50	12843.10		
Y4	1		5594.70	302.96	5.42	5000.90	6188.50		
Y5	0		17514.00	93.42	0.53	17330.89	17697.11		
Y5	1		330.00	93.42	28.31	146.89	513.11		
Y6	0		15625.60	245.29	1.57	15144.82	16106.38		
Y6	1		2218.40	245.29	11.06	1737.62	2699.18		

dominio pianificato=PROV1 sottoclasse=SEX modalità di sottoclasse=2									
variabili di interesse	variabili di interesse	modalità di interesse	stima	errore standard	errore relativo %	limite inf. I.C.	limite sup. I.C.		
Y1	0		12386.00	289.78	2.34	11818.03	12953.97		
Y1	1		6428.00	289.78	4.51	5860.03	6995.97		
Y2	0		12882.90	293.36	2.28	12307.91	13457.89		
Y2	1		5931.10	293.36	4.95	5356.11	6506.09		
Y3	0		17607.80	205.56	1.17	17204.90	18010.70		
Y3	1		1206.20	205.56	17.04	803.30	1609.10		
Y4	0		14089.10	292.13	2.07	13516.53	14661.67		
Y4	1		4724.90	292.13	6.18	4152.33	5297.47		
Y5	0		18746.50	36.28	0.19	18675.39	18817.61		
Y5	1		67.50	36.28	53.75	-3.61	138.61		
Y6	0		18311.90	117.33	0.64	18081.94	18541.86		
Y6	1		502.10	117.33	23.37	272.14	732.06		

Commenti relativi alla stampa 1:

La stima del totale di occupati per l'intero territorio nazionale è pari a 197784, con errore relativo percentuale pari a 1,9% (vedasi riga 4 del dominio pianificato=TOTALE, sottoclasse=0, modalità di sottoclasse=0).

Analoghe informazioni possono essere desunte per ciascuna provincia; si

veda ad esempio la provincia = prov1 (dominio pianificato=PROV1, sottoclasse=0, modalità di sottoclasse=0).

Le informazioni relative al totale di occupati distinti per sesso possono essere ricavate facendo riferimento alla riga 4 della stampa di dominio pianificato=TOTALE, sottoclasse=SEX, modalità di sottoclasse=1 o 2, per i maschi o per le femmine (secondo la codifica adottata nel data set di input).

Stampe 2, 3, 4

Nelle stampe 2, 3 e 4 sono contenute alcune informazioni sul disegno di campionamento per dominio di stima. Analogamente alla stampa 1, le informazioni sono organizzate secondo il dominio pianificato, la sottoclasse e la modalità di sottoclasse. Le stampe 3 e 4 contengono le informazioni relative alla parte di campione in cui il “Tipo di disegno” è “1” o “0”, che nell’applicazione permettono di distinguere le informazioni dei comuni autorappresentativi da quelle dei comuni non autorappresentativi.

A titolo esemplificativo viene riportato solo il caso relativo all’intero territorio nazionale.

(da stampa2.txt, stampa3.txt)

2 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative

dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0

variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	numero unità elementari	stima del totale unità ementari
Y1	0	0.491	0.76	0.17	3000	534755
Y1	1	0.491	0.76	0.22	3000	534755
Y2	0	0.483	0.80	0.17	3000	534755
Y2	1	0.483	0.80	0.25	3000	534755
Y3	0	0.320	1.23	0.12	3000	534755
Y3	1	0.320	1.23	0.60	3000	534755
Y4	0	0.435	1.19	0.18	3000	534755
Y4	1	0.435	1.19	0.45	3000	534755
Y5	0	0.156	1.44	0.06	3000	534755
Y5	1	0.156	1.44	0.90	3000	534755
Y6	0	0.329	1.45	0.15	3000	534755
Y6	1	0.329	1.45	0.53	3000	534755

3 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative - tipo di disegno=1

dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0

variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	correlaz. intraclasse	numero unità elementari
Y1	0	0.491	0.94	0.58	-0.076	1849
Y1	1	0.491	0.94	0.65	-0.076	1849
Y2	0	0.484	0.96	0.57	-0.053	1849
Y2	1	0.484	0.96	0.67	-0.053	1849
Y3	0	0.310	1.52	0.49	0.863	1849
Y3	1	0.310	1.52	0.96	0.863	1849
Y4	0	0.442	1.16	0.58	0.222	1849
Y4	1	0.442	1.16	0.79	0.222	1849
Y5	0	0.066	1.43	0.10	0.686	1849
Y5	1	0.066	1.43	1.19	0.686	1849
Y6	0	0.279	1.47	0.44	0.749	1849
Y6	1	0.279	1.47	0.94	0.749	1849

(da stampa4.txt)

4 - Informazioni sul disegno di campionamento per dominio di stima variabili qualitative - tipo disegno=0						
dominio pianificato=TOTALE sottoclasse=0 modalità di sottoclasse=0						
variabili di interesse	modalità variabili interesse	scarto q. medio	deft	effetto stimatore	correlaz. intraclasse	numero unità elementari
Y1	0	0.491	0.58	0.11	-0.008	1151
Y1	1	0.491	0.58	0.14	-0.008	1151
Y2	0	0.482	0.65	0.11	-0.007	1151
Y2	1	0.482	0.65	0.17	-0.007	1151
Y3	0	0.329	0.96	0.08	-0.001	1151
Y3	1	0.329	0.96	0.45	-0.001	1151
Y4	0	0.429	1.15	0.14	0.004	1151
Y4	1	0.429	1.15	0.38	0.004	1151
Y5	0	0.208	1.27	0.06	0.007	1151
Y5	1	0.208	1.27	0.89	0.007	1151
Y6	0	0.367	1.34	0.13	0.009	1151
Y6	1	0.367	1.34	0.47	0.009	1151

Commenti alle stampe 2,3 e 4:

Per la variabile disoccupati (Y3), si desume che l'effetto del disegno e dello stimatore (per la definizione cfr. *paragrafo 5.3, Sezione I*) adottati è pari a 1,52 per la parte del campione autorappresentativa (tipo disegno pari a 1; *stampa 3*) mentre è pari a 0,96 per la parte del campione non autorappresentativa (tipo disegno pari a 0; *stampa 4*). Ciò implica che la strategia adottata è inaspettatamente più efficiente per la componente non autorappresentativa. L'incoerenza del risultato può, tuttavia, essere spiegata dalla natura fittizia dei dati.

La lettura delle informazioni contenute nella *stampa 2* permette di valutare la strategia di campionamento nel suo complesso, prescindendo quindi dalla suddivisione dei comuni in componente autorappresentativa e non autorappresentativa. Ad esempio, l'effetto complessivo dello stimatore di calibrazione utilizzato per la variabile numero di disoccupati è pari a 0,60.

La correlazione intraclasse è riportata solo nelle *stampe 3 e 4*, essendo priva di significato a livello complessivo, l'unità primaria è, infatti, definita in modo differente nella parte autorappresentativa e in quella non autorappresentativa.

Stampa 6

La stampa 6 presenta, per ciascun dominio di stima, alcune informazioni sintetiche sul disegno di campionamento. A titolo di esempio, viene riportato il risultato di tale stampa solo per la provincia PROV3.

(da stampa6.txt)

6-Informazioni sintetiche sul disegno di campionamento per dominio di stima						
dominio pianificato=PROV3 sottoclasse=0						
modalità sottoclasse	deft medio	deft massimo	effetto stim. medio	effetto stim. massimo	errore rel. % medio	errore rel. % massimo
0	0.95	1.55	0.25	0.96	6.4	24.5
dominio pianificato=PROV3 sottoclasse=SEX						
modalità sottoclasse	deft medio	deft massimo	effetto stim. medio	effetto stim. massimo	errore rel. % medio	errore rel. % massimo
1	0.64	1.33	0.29	1.08	7.9	36.0
2	0.78	1.58	6.09	69.73	10.8	27.8

Commenti relativi alla stampa 6:

Si può osservare che nella provincia PROV3, il valore massimo dell'errore relativo percentuale nell'insieme delle 6 variabili è pari a 24,5%, il valore medio dell'errore relativo per i maschi è, invece, pari a 7,9%, mentre per le femmine è pari a 10,8%.

Stampe 5b, 5b, 7a, 7b

Le stampe 5a, 5b, 7a e 7b sono relative alla presentazione sintetica degli errori di campionamento. Poiché ad ogni stima campionaria \hat{Y} corrisponde un errore di campionamento relativo $\hat{\epsilon}(\hat{Y})$, nelle tabelle pubblicate si dovrebbe associare ad ogni stima il corrispondente errore di campionamento relativo. Tuttavia per limiti di tempo, per costi di elaborazione, e per facilitare la consultazione delle tavole, si preferisce omettere tale informazione, che non sarebbe comunque disponibile per le stime che l'utente decide di ricavare autonomamente. Si preferisce, quindi, dare una presentazione sintetica degli errori relativi, basata sul *metodo dei modelli regressivi*. Il software permette di adattare con il metodo dei minimi quadrati due differenti modelli i cui risultati sono riportati nelle stampe 5a e 7a.

La stampa 5a presenta i valori dei coefficienti $\hat{\alpha}_1$, $\hat{\alpha}_2$ e dell'indice di

determinazione R^2 del primo modello $\log(\hat{\epsilon}^2(\hat{Y})) = \hat{\alpha}_1 + \hat{\alpha}_2 \log(\hat{Y})$, generalmente applicato per le stime di frequenze assolute e relative. I parame-

tri A e B della stampa 5a corrispondono ad \hat{e} e $\hat{\sigma}^2$ del modello di cui sopra. A, B ed R^2 sono stimati a livello di totale Italia e di ciascuna provincia.

(da stampa5.txt)

5a - Valori dei parametri A e B e indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari			
dominio pianificato	A	B	indice di determinazione
PROV1	7.2217	-1.59775	81.67
PROV2	11.0302	-1.64744	83.83
PROV3	11.3576	-1.70501	71.03
PROV4	10.9205	-1.68725	84.88
PROV5	11.9289	-1.76303	89.37
TOTALE	12.7455	-1.72878	88.30

Commenti relativi alla stampa 5a:

Sostituendo nel modello le stime ottenute è possibile dare, anche per grandezze non riportate nelle stampe precedenti, una valutazione approssimativa dell'errore. Ad esempio, se si considera una stima pari a 250.000 per la provincia PROV5, utilizzando i dati della stampa 5a, si ottiene che : $\log(\hat{\sigma}^2(\hat{Y})) = 11,9289 - 1,7630 \log(250000) = 2,4122$ e pertanto $\hat{e}(\hat{Y}) = 3,3404$.

La valutazione approssimata degli errori è resa ancora più agevole dalla stampa 5b, dove sono riportati, accanto ad alcune grandezze di riferimento, i rispettivi valori degli errori calcolati secondo il modello.

(da stampa5.txt)

5b - Valori interpolati degli errori di campionamento per dominio di stima pianificato						
stima %	PROV1		PROV2		PROV3	
	stima	errore rel.%	stima	errore rel.%	stima	errore rel.%
0.10	36.66	208.26	140.42	422.96	122.75	484.60
0.50	183.29	57.57	702.09	112.34	613.74	122.89
1.00	366.58	33.09	1404.19	63.47	1227.48	68.06
2.00	733.16	19.02	2808.37	35.86	2454.96	37.69
3.00	1099.74	13.76	4212.56	25.68	3682.44	26.68
4.00	1466.32	10.93	5616.74	20.26	4909.92	20.87
5.00	1832.90	9.15	7020.93	16.86	6137.41	17.26
6.00	2199.48	7.91	8425.11	14.51	7364.89	14.77
7.00	2566.06	6.99	9829.30	12.78	8592.37	12.95
8.00	2932.64	6.28	11233.48	11.45	9819.85	11.56
9.00	3299.22	5.72	12637.67	10.39	11047.33	10.46
10.00	3665.80	5.26	14041.85	9.52	12274.81	9.56
15.00	5498.70	3.80	21062.78	6.82	18412.22	6.76
20.00	7331.60	3.02	28083.70	5.38	24549.62	5.29
25.00	9164.50	2.53	35104.63	4.48	30687.03	4.38
30.00	10997.40	2.19	42125.55	3.85	36824.43	3.75
35.00	12830.30	1.93	49146.48	3.39	42961.84	3.29
40.00	14663.20	1.74	56167.40	3.04	49099.24	2.93
45.00	16496.10	1.58	63188.33	2.76	55236.65	2.65
50.00	18329.00	1.45	70209.25	2.53	61374.05	2.42

Commenti alla stampa 5b:

Nella stampa 5b sono riportate, in valore percentuale, alcune grandezze di riferimento e i rispettivi errori, calcolati in base al modello adottato. Ad esempio nella provincia PROV1, ad una stima pari a 12% si può associare un errore relativo pari a 5,26% corrispondente al valore tabulato pari a 10%, e che quindi rappresenta un valore conservativo dell'errore.

Le stampe 7a e 7b si riferiscono, invece, al modello regressivo

$$\hat{\epsilon}(\hat{Y}) = \hat{\alpha}_2 + \frac{\hat{\alpha}_1}{\hat{Y}} + \hat{\alpha}_3 \hat{Y}$$
 e danno informazioni analoghe a quelle delle

stampe 5a e 5b. . I parametri A, B e C della stampa 7a corrispondono ad

$\hat{\alpha}_1$, $\hat{\alpha}_2$ e $\hat{\alpha}_3$ del modello di cui sopra.

Nell' *appendice A.5* sono descritti dettagliatamente i precedenti modelli e il significato delle relative stampe.

(da stampa7.txt)

7a - Modello alternativo				
valori dei parametri e indice di determinazione per dominio di stima pianificato del modello di regressione per la presentazione sintetica degli errori campionari				
dominio pianificato	A	B	C	indice di determinazione
PROV1	180.91	0.011325	-.000000281	12.94
PROV2	1380.27	0.011084	-.000000064	6.54
PROV3	1293.89	0.008716	-.000000039	3.03
PROV4	955.54	0.012160	-.000000079	10.40
PROV5	1228.29	0.002742	0.000000018	8.97
TOTALE	2573.09	0.003556	-.000000003	7.74

(da stampa7.txt)

7b - Modello alternativo Valori interpolati degli errori di campionamento per dominio di stima pianificato						
	PROV1		PROV2		PROV3	
stima %	stima	errore rel. %	stima	errore rel. %	stima	errore rel. %
0.01	3.67	4936.34	14.04	9830.82	12.27	10541.92
0.02	7.33	2468.74	28.08	4915.96	24.55	5271.40
0.03	11.00	1646.20	42.13	3277.68	36.82	3514.55
0.04	14.66	1234.93	56.17	2458.53	49.10	2636.13
0.05	18.33	988.17	70.21	1967.05	61.37	2109.08
0.10	36.66	494.65	140.42	984.08	122.75	1054.98
0.50	183.29	99.83	702.09	197.70	613.74	211.69
1.00	366.58	50.47	1404.19	99.40	1227.48	106.28
2.00	733.16	25.79	2808.37	50.24	2454.96	53.57
3.00	1099.74	17.55	4212.56	33.85	3682.44	35.99
4.00	1466.32	13.43	5616.74	25.65	4909.92	27.20
5.00	1832.90	10.95	7020.93	20.72	6137.41	21.93
10.00	3665.80	5.96	14041.85	10.85	12274.81	11.36
15.00	5498.70	4.27	21062.78	7.53	18412.22	7.83
20.00	7331.60	3.39	28083.70	5.84	24549.62	6.05
25.00	9164.50	2.85	35104.63	4.81	30687.03	4.97
30.00	10997.40	2.47	42125.55	4.11	36824.43	4.24
35.00	12830.30	2.18	49146.48	3.60	42961.84	3.71
40.00	14663.20	1.95	56167.40	3.21	49099.24	3.31
45.00	16496.10	1.77	63188.33	2.89	55236.65	3.00
50.00	18329.00	1.60	70209.25	2.62	61374.05	2.74

Stampa 8

La stampa 8, infine, contiene informazioni sul processo di aggregazione degli strati e permette di approfondire l'analisi di controllo relativa alla suddetta aggregazione. Per ogni strato originario si può verificare se è stata effettuata un'aggregazione e a quale strato finale del processo è associata. Il valore di "tipo aggreg", che appare nella prima colonna della stampa, è 0 se lo strato non è da aggregare; 1 e 2 se è da aggregare perché costituito da una sola unità primaria. In particolare, assume valore pari ad 1 se è possibile aggregare tale strato, assume valore pari a 2 se non è possibile effettuare tale aggregazione.

(da stampa8.txt)

8 - Analisi di controllo sulla aggregazione degli strati: caso in cui i superstrati sono formati da 2 strati originari

Popolaz. pianif. utiliz. per stimatore=REG1 dominio pianificato=PROV1

	codice tipo strato aggreg. orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	numero tipo disegno	stima totale unità finali	stima totale unità elem.
	0 str01	1	399	399	964	1	14763	35668
	2 str02	2	1	15	36	0	555	1332
-----			-----		-----		-----	-----
dominio_pianificato				414	1000		15318	37000
popolaz_pianificata				414	1000		15318	37000

Popolaz. pianif. utiliz. per stimatore=REG2 dominio pianificato=PROV2

	codice tipo strato aggreg. orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	numero tipo disegno	stima totale unità finali	stima totale unità elem.
	0 str03	1	96	96	216	1	30144	67824
	0 str04	2	56	56	146	1	14672	38252
	1 str05	3	1	12	32	0	3348	8928
	1 str06	3	1	40	106	0	10600	28090
-----			-----		-----		-----	-----
dominio_pianificato				204	500		58764	143094

Popolaz. pianif. utiliz. per stimatore=REG2 dominio pianificato=PROV3

	codice tipo strato aggreg. orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	numero tipo disegno	stima totale unità finali	stima totale unità elem.
	0 str08	4	2	86	238	0	21670	59972
	0 str09	5	2	47	120	0	12292	31477
	2 str07	6	1	40	142	0	7760	27548
-----			-----		-----		-----	-----
dominio_pianificato				173	500		41722	118997

Popolaz. pianif. utiliz. per stimatore=REG2 dominio pianificato=PROV4

	codice tipo strato aggreg. orig.	codice supstr. finale	numero unità primarie	numero unità finali	numero unità elem.	numero tipo disegno	stima totale unità finali	stima totale unità elem.
	0 str10	7	47	47	123	1	12361	32349
	0 str11	8	52	52	156	1	12272	36816
	1 str12	9	1	32	99	0	7200	22275

(segue da stampa8.txt)

1 str13	9	1	32	122	0	5856	22326
-----			-----		-----		
dominio_pianificato			163	500		37689	113766
Popolaz. pianif. utiliz. per stimatore=REG2 dominio pianificato=PROV5							
codice	codice	numero	numero	numero		stima	stima
tipo strato	supstr.	unità	unità	unità	tipo	totale	totale
aggreg. orig.	finale	primarie	finali	elem.	disegno	finali	elem.
0 str14	10	80	80	244	1	18400	56120
1 str15	11	1	36	100	0	9360	26000
1 str16	11	1	36	105	0	9180	26775
1 str17	11	1	19	51	0	5168	13872
-----			-----		-----		
dominio_pianificato			171	500		42108	122767
popolaz_pianificata			711	2000		180283	498624
			=====		=====		
			1125	3000		195601	535624

Commenti alla stampa 8:

Come già è stato osservato, nel caso in esame gli strati str02 e str07 non sono aggregabili e presentano codice 2. Si noti che gli strati str15, str16 e str17 della PROV5 risultano comporre il superstrato finale numero 11. In questo caso poiché gli strati originari che presentano una sola unità primaria sono tre, lo strato finale risulta formato da più di due (parametro di *collassamento*) strati iniziali.

APPENDICI

A.1 Cenni sulla definizione dello stimatore di regressione generalizzata

Per descrivere la metodologia adottata dal software generalizzato per il calcolo degli errori di campionamento per la stima di un totale, si prenda in considerazione una popolazione $U = \{1, \dots, k, \dots, N\}$, di N elementi, e si denoti con Y la variabile oggetto d'indagine.

Sia quindi:

$$Y = \sum_{k \in U} y_k$$

il parametro da stimare, essendo y_k il valore della variabile d'interesse Y assunto dalla generica unità k .

Il software permette di calcolare gli errori campionari di un'ampia classe di stimatori diretti di Y , i quali possono essere derivati dalla teoria degli stimatori di regressione generalizzata. Tali stimatori appartengono, a loro volta, alla classe degli stimatori di calibrazione che, in estrema sintesi, definiscono i coefficienti finali delle unità attraverso la risoluzione di un problema di minimo vincolato. In particolare, dati dei totali noti a livello di popolazione (o sottopopolazione), per alcune variabili ausiliarie il processo di ottimizzazione avviene minimizzando la distanza tra i coefficienti diretti (pari all'inverso della probabilità di inclusione nel campione), eventualmente corretti in presenza di mancate risposte totali, e i coefficienti finali (incogniti) assegnati alle unità campionarie, con il vincolo che le stime ottenute con i coefficienti finali riproducano i totali noti sopra definiti.

Ciascuno stimatore di calibrazione si distingue sia per il tipo di totali noti utilizzati che per altri due elementi riguardanti: la funzione di distanza impiegata, per valutare lo scostamento tra i coefficienti diretti e quelli

finali; il peso, c_k , attribuito a ciascuna unità del campione, che interviene come fattore moltiplicativo della distanza calcolata tra coefficiente diretto e finale per l'unità k -esima.

Si dimostra che gli stimatori di regressione generalizzata sono un caso particolare degli stimatori di calibrazione, quando la distanza scelta per l'ottenimento dei pesi finali è quella euclidea (Deville e Särndal 1992). In tal caso, con riferimento ad un campione casuale $s = \{1, \dots, k, \dots, n\}$ di n unità, il problema di minimo vincolato è rappresentato dal seguente sistema

$$\begin{cases} \min \left[\sum_{k \in s} \frac{(\frac{1}{\pi_k} - w_k)^2}{\frac{1}{\pi_k}} \cdot c_k \right], \\ \sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X} \end{cases}$$

in cui, relativamente alla k -esima unità appartenente al campione, si ha che nella prima espressione (funzione obiettivo) π_k è la probabilità di inclusione, w_k è il peso finale calibrato incognito e c_k è un peso indipendente da π_k attribuito a ciascuna unità del campione.

Nella seconda espressione, detta *equazione di calibrazione*, sono contenuti i vincoli $\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$ e rappresenta il vettore dei valori assunti dalle J variabili ausiliarie $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)$ per le quali sono noti i totali $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)'$ riferiti all'intera popolazione (o eventualmente a particolari sottopopolazioni).

Un importante risultato ottenuto in Deville e Särndal (1992) indica che, nelle indagini su larga scala, gli stimatori di calibrazione che utilizzano una generica funzione di distanza sono asintoticamente equivalenti ai corrispondenti stimatori di regressione generalizzata che usano la distanza euclidea¹⁰. Alla luce di questo risultato la stima della varianza di tutti gli stimatori di calibrazione può essere approssimata dalla stima della varianza calcolata sui corrispondenti stimatori di regressione per i quali è possibile derivare l'espressione esplicita della stima della varianza.

¹⁰

Più precisamente per assicurare l'equivalenza asintotica fra le stime prodotte con uno stimatore di calibrazione e quelle prodotte con uno stimatore di regressione generalizzata, la funzione di distanza del primo stimatore deve rispettare alcune deboli condizioni (Deville e Särndal 1992).

Restringendo pertanto l'attenzione alla classe degli stimatori di regressione generalizzata, secondo una trattazione generale questi si fondano sulle seguenti informazioni:

- per ciascun elemento del campione k si conosce il vettore delle $J+1$ osservazioni (y_k, \mathbf{x}_k) , in cui $\mathbf{x}_k = (x_{1k}, \dots, x_{jk}, \dots, x_{Jk})'$ è il vettore dei valori assunti dalle J variabili ausiliarie $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)$;
- risulta noto il vettore dei totali $\mathbf{X} = (X_1, \dots, X_j, \dots, X_J)'$ corrispondenti alle J variabili ausiliarie.

Lo stimatore in questione sfrutta le suddette informazioni ausiliarie attraverso la definizione di un modello di regressione lineare ξ che spiega la nuvola dei punti individuata dall'insieme $\{(y_k, \mathbf{x}_k) : k=1, \dots, N\}$. Il modello si basa sulle seguenti ipotesi:

- i valori $y_1, \dots, y_k, \dots, y_N$ assunti dalla variabile Y per le N unità della popolazione sono considerati come realizzazioni di N variabili casuali indipendenti;
- le variabili ausiliarie sono trattate come costanti note di tipo non stocastico;
- la relazione che lega la generica variabile casuale y_k al vettore \mathbf{x}_k ($k=1, \dots, N$) è la seguente:

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k, \quad (k=1, \dots, N) \quad (\text{A.1.1})$$

in cui $\boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_J)'$ è il vettore dei J coefficienti di regressione incogniti ed ε_k è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello ξ sono definiti rispettivamente da:

$$E_\xi(\varepsilon_k) = 0, \quad \text{Var}_\xi(\varepsilon_k) = c_k \sigma^2, \quad \text{Cov}_\xi(\varepsilon_k, \varepsilon_l) = 0 \quad \text{per } \forall k \neq l; \quad (\text{A.1.2})$$

essendo c_k (per $k \in U$) delle costanti note.

Si supponga di aver effettuato un censimento di tutte le N unità della popolazione U e di disporre, quindi, di tutti i valori della nuvola di punti. E' possibile utilizzare, allora, la nuvola di punti della popolazione per stimare, mediante il metodo dei minimi quadrati ponderati, il vettore dei coefficienti di regressione $\boldsymbol{\beta}$ del modello ξ . Utilizzando la teoria standard

della regressione generalizzata, si ha che il miglior stimatore lineare non distorto dei coefficienti \mathbf{B} , sotto il modello ξ , è dato da:

$$\mathbf{B} = (B_1, \dots, B_j, \dots, B_J)' = \left(\sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \right)^{-1} \sum_{k \in U} \frac{\mathbf{x}_k y_k}{c_k} . \quad (\text{A.1.3})$$

Il vettore dei coefficienti \mathbf{B} è, tuttavia, una caratteristica incognita della popolazione in quanto le variabili \mathbf{X} e Y non sono note per l'intero universo. Si può, pertanto, procedere ad una stima di \mathbf{B} mediante i dati rilevati sul campione s . Poiché la relazione (A.1.3) si presenta come il prodotto di una funzione dei totali della popolazione;

$$\mathbf{T}_1 = \sum_{k \in U} \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k} \quad \text{e} \quad \mathbf{T}_2 = \sum_{k \in U} \frac{\mathbf{x}_k y_k}{c_k} ,$$

una stima asintoticamente corretta di \mathbf{B} può essere ottenuta stimando ciascun totale mediante lo stimatore di Horvitz-Thompson. I due stimatori sono espressi attraverso le seguenti formule

$$\hat{\mathbf{T}}_1 = \sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k} \quad \text{e} \quad \hat{\mathbf{T}}_2 = \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\pi_k c_k} .$$

La stima di \mathbf{B} assume, pertanto, la seguente forma:

$$\hat{\mathbf{B}} = (\hat{B}_1, \dots, \hat{B}_j, \dots, \hat{B}_J)' = \hat{\mathbf{T}}_1^{-1} \hat{\mathbf{T}}_2 = \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k} \right)^{-1} \sum_{k \in s} \frac{\mathbf{x}_k y_k}{\pi_k c_k}$$

Sulla base di $\hat{\mathbf{B}}$ è possibile, quindi, calcolare con riferimento alle N unità della popolazione, i valori interpolati $\hat{y}_1, \dots, \hat{y}_k, \dots, \hat{y}_N$, relativi ai corrispondenti valori $y_1, \dots, y_k, \dots, y_N$, mediante la relazione

$$\hat{y}_k = \mathbf{x}_k' \hat{\mathbf{B}}, \quad (k = 1, \dots, N) . \quad (\text{A.1.4})$$

E' opportuno sottolineare che questa versione del software utilizza per la stima di \mathbf{B} i coefficienti finali di riporto presenti nel *data-set* di input. Tale accorgimento conduce a stime più efficienti della varianza.

Inoltre, con riferimento alle n unità del campione e in base alla (A.1.4) i residui sono dati da

$$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}, \quad (k = 1, \dots, n) . \quad (\text{A.1.5})$$

Per la (A.1.5), il totale Y può, quindi, essere riscritto mediante la seguente espressione

$$Y = \sum_{k \in U} y_k = \sum_{k \in U} \hat{y}_k + \sum_{k \in U} e_k . \quad (\text{A.1.6})$$

Dalla (A.1.6) si osserva che l'ultima relazione dopo il segno di uguaglianza è costituita dalla somma di due totali: il primo è una quantità nota, in quanto il \hat{y}_k valore può essere definito per tutte le unità della popolazione; il secondo, invece, rappresenta una quantità incognita, poiché è possibile calcolare i residui solo per le unità appartenenti al campione osservato. Sostituendo quindi nella (A.1.6) lo stimatore di Horvitz-Thompson di tale totale incognito, si ottiene lo stimatore di regressione generalizzata del totale Y

$$\hat{Y}_{\text{GREG}} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} \frac{e_k}{\pi_k} . \quad (\text{A.1.7})$$

Considerando che il termine $\sum_{k \in U} \hat{y}_k$ si può riformulare come

$$\sum_{k \in U} \hat{y}_k = \sum_{k \in U} \mathbf{x}'_k \hat{\mathbf{B}} = \left(\sum_{k \in U} \mathbf{x}_k \right)' \hat{\mathbf{B}} = \mathbf{X}' \hat{\mathbf{B}} \quad (\text{A.1.8})$$

e che il secondo totale delle (A.1.7) può essere riscritto mediante il seguente passaggio

$$\sum_{k \in s} \frac{e_k}{\pi_k} = \sum_{k \in s} \frac{(y_k - \mathbf{x}'_k \hat{\mathbf{B}})}{\pi_k} = \sum_{k \in s} \left(\frac{y_k}{\pi_k} \right) - \sum_{k \in s} \left(\frac{\mathbf{x}_k}{\pi_k} \right)' \hat{\mathbf{B}} = \hat{Y} - \hat{\mathbf{X}}' \hat{\mathbf{B}} , \quad (\text{A.1.9})$$

in cui \hat{Y} e $\hat{\mathbf{X}}$ indicano le stime di Horvitz-Thompson dei corrispondenti totali Y e \mathbf{X} , è possibile riformulare la (A.1.7) secondo l'espressione

$$\hat{Y}_{\text{GREG}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{B}} \quad (\text{A.1.10})$$

dalla quale risulta che lo stimatore di regressione generalizzata è ottenuto come somma dello stimatore di Horvitz-Thompson del totale Y più un termine di aggiustamento regressivo che dipende dalle differenze tra i totali noti \mathbf{X} e le corrispondenti stime campionarie di Horvitz-Thompson $\hat{\mathbf{X}}$ ponderate con i rispettivi coefficienti di regressione stimati $\hat{\mathbf{B}}$.

Dalla (A.1.10), attraverso alcuni semplici passaggi lo stimatore si può riscrivere come

$$\hat{Y}_{GREG} = \sum_{k \in s} \frac{g_{ks} y_k}{\pi_k} . \quad (\text{A.1.11})$$

dove compare il fattore correttivo del peso diretto $1/\pi_k$:

$$g_{ks} = 1 + (X - \hat{X})' \left(\sum_{k \in s} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k c_k} \right)^{-1} \frac{\mathbf{x}_k}{c_k} . \quad (\text{A.1.12})$$

Una importante proprietà dello stimatore di regressione generalizzata è che la stima dei totali di popolazione delle variabili ausiliarie è uguale ai corrispondenti totali noti. Sostituendo nella (A.1.11) y_k con \mathbf{x}_k si ha, infatti,

$$\sum_{k \in s} \frac{g_{ks} \mathbf{x}_k}{\pi_k} = X .$$

Una definizione più precisa dello stimatore di regressione generalizzata passa attraverso l'introduzione di tre concetti che specificano ulteriormente la relazione della variabile d'interesse con il relativo modello di regressione. Questi sono: il *gruppo di riferimento del modello* (*model group*), il *livello del modello* (*model level*) ed il *tipo di modello* (*model type*).

A.1.1 Gruppo di riferimento del modello

Data una partizione completa della popolazione $U, \{U_1, \dots, U_d, \dots, U_D\}$, si definisce il generico *gruppo di riferimento del modello* U_d un sottoinsieme (o sottopopolazione) in cui:

- sono noti i totali di una o più variabili ausiliarie. Occorre notare che non è necessario che l'insieme delle variabili ausiliarie sia lo stesso per ciascuna sottopopolazione.
- il campione s_d appartenente al gruppo di riferimento d , definito come, $s_d = s \cap U_d$ deve essere sempre costituito da un numero di unità maggiore del numero di totali noti.

Valendo le precedenti condizioni è possibile definire un modello separato per le unità di ciascun gruppo. Rispetto alla (A.1.1), in cui il gruppo di riferimento è l'intero universo U , si costruisce quindi un modello di regressione per ciascun U_d , espresso da

$$y_k = \mathbf{x}_{dk}' \boldsymbol{\beta}_d + \varepsilon_k \quad \forall k \in U_d , \quad (\text{A.1.13})$$

in cui valgono le ipotesi (A.1.2) ed in cui \mathbf{x}_{dk} è il vettore dei valori assunti, dall'unità k , sulle variabili ausiliarie utilizzate per la costruzione del modello, nella sottopopolazione U_d .

Analogamente alla (A.1.3) la stima del vettore β_d si ottiene come:

$$\hat{\mathbf{B}}_d = \left(\sum_{k \in s_d} \frac{\mathbf{x}_{dk} \mathbf{x}'_{dk}}{\pi_k c_k} \right)^{-1} \sum_{k \in s_d} \frac{\mathbf{x}_{dk} y_k}{\pi_k c_k}.$$

Lo stimatore di regressione generalizzata basato su una suddivisione dell'universo in gruppi di riferimento è dato da:

$$\hat{Y}_{GREG} = \sum_{d=1}^D \sum_{k \in s_d} \frac{g_{ks_d} y_k}{\pi_k},$$

nella quale

$$g_{ks_d} = 1 + (\mathbf{X}_d - \hat{\mathbf{X}}_d)' \left(\sum_{k \in s_d} \frac{\mathbf{x}_{dk} \mathbf{x}'_{dk}}{\pi_k c_k} \right)^{-1} \frac{\mathbf{x}_{dk}}{c_k} \quad (\text{A.1.14})$$

$$\text{con } \mathbf{X}_d = \sum_{U_d} \mathbf{x}_{dk} \quad \text{e} \quad \hat{\mathbf{X}}_d = \sum_{s_d} \mathbf{x}_{dk} / \pi_k.$$

A.1.2 Livello del modello

Il concetto di livello del modello è relativo al tipo di unità utilizzata nella formulazione del modello. Ad esempio il modello può essere formulato a livello di:

- a) *unità elementare*, se nella sua definizione le variabili d'interesse e quelle ausiliarie si riferiscono a ciascuna unità elementare della popolazione;
- b) *cluster* (o gruppi) di elementi, se nella sua definizione le variabili d'interesse e quelle ausiliarie si riferiscono a grappoli di unità elementari della popolazione.

In assenza di gruppi di riferimento del modello il caso a) prevede che nella relazione (A.1.1) e sotto le ipotesi (A.1.2), k indichi la generica unità elementare.

Per il caso b), definito con $U_I = \{1, \dots, i, \dots, N_I\}$, l'universo dei *cluster*, si può costruire il seguente modello di regressione ξ_I

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta}_I + \varepsilon_i \quad , \quad (i=1, \dots, N_I) \quad (\text{A.1.15})$$

in cui $Y_i = \sum_{k \in i} y_k$ e $\mathbf{X}_i = \sum_{k \in i} \mathbf{x}_k$ sono i totali di Y e \mathbf{X} per il generico cluster i ;

$\boldsymbol{\beta}_I = (\beta_{I1}, \dots, \beta_{Ij}, \dots, \beta_{IJ})'$ è il vettore dei J coefficienti di regressione incogniti;

ε_i è una variabile casuale per la quale il valore atteso, la varianza e la covarianza sotto il modello ξ_I sono definiti rispettivamente da:

$$E_{\xi_I}(\varepsilon_i) = 0, \quad Var_{\xi_I}(\varepsilon_i) = c_i \sigma_I^2, \quad Cov_{\xi_I}(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{per } \forall i \neq i'; \quad (\text{A.1.16})$$

essendo c_i (per $i \in U_I$) le delle costanti note.

Lo stimatore di regressione definito a livello di *cluster* dato dalla (A.1.15) e dalla (A.1.16) assume, dunque, la seguente espressione:

$$\hat{Y}_{GREG} = \sum_{i \in s_I} \frac{g_{ks_I} Y_i}{\pi_i}$$

in cui s_I è il campione dei cluster e

$$g_{ks_I} = 1 + (\mathbf{X} - \hat{\mathbf{X}})' \left(\sum_{i \in s_I} \frac{\mathbf{X}_i \mathbf{X}_i'}{\pi_i c_i} \right)^{-1} \frac{\mathbf{X}_i}{c_i} \quad (\text{A.1.17})$$

è il fattore correttivo del peso diretto e π_i è la probabilità di inclusione del cluster i nel campione s_I .

Espressioni analoghe alla (A.1.13) e alla (A.1.14) si ottengono quando la popolazione U_I è partizionata in $U_{I1}, \dots, U_{Id}, \dots, U_{ID}$ gruppi di riferimento. La relazione che lega la variabile oggetto d'indagine e le variabili ausiliarie è data da

$$Y_i = \mathbf{X}_{di}' \boldsymbol{\beta}_{Id} + \varepsilon_i \quad \forall i \in U_{Id}$$

in cui \mathbf{X}_{di} è il vettore dei totali calcolati sul cluster i delle variabili ausiliarie utilizzate per la costruzione del modello nella sottopopolazione U_{Id} .

Lo stimatore di regressione si può, pertanto, formulare attraverso la relazione

$$\hat{Y}_{GREG} = \sum_{d=1}^D \sum_{k \in s_{Id}} \frac{g_{ks_{Id}} y_k}{\pi_k} \quad ,$$

in cui $s_{Id} = s_I \cap U_{Id}$;

$$g_{ks_{Id}} = 1 + (\mathbf{X}_d - \hat{\mathbf{X}}_d)' \left(\sum_{i \in s_{Id}} \frac{\mathbf{X}_{di} \mathbf{X}_{di}'}{\pi_i c_i} \right)^{-1} \frac{\mathbf{X}_{di}}{c_i} \quad (\text{A.1.18})$$

è il fattore correttivo calcolato a livello di *cluster*.

Si ricorda che un modello a livello di unità elementare corrisponde ad uno stimatore che attribuisce un peso finale diverso per tutte le unità elementari appartenenti ad una medesima unità finale di campionamento; viceversa, un modello a livello di *cluster* di unità elementari corrisponde ad uno stimatore che attribuisce un peso finale uguale per tutte le unità elementari appartenenti ad una medesima unità finale di campionamento.

Infine si ricorda che, mentre per impostare un modello a livello di unità elementare non vi sono vincoli sul tipo di disegno campionario adottato, per definire nel software un modello di regressione a livello di cluster è necessario aver utilizzato un disegno in cui le unità finali di campionamento sono dei grappoli.

A.1.3 Tipo di modello

La scelta delle variabili ausiliarie $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_j, \dots, \mathbf{X}_J)$ e del parametro c_k determina il *tipo di modello* sottostante allo stimatore di regressione generalizzata. In particolare, la specificazione di \mathbf{X} e c_k , associata alla definizione del livello e del gruppo di riferimento, conducono a noti stimatori che possono essere derivati anche al di fuori della teoria degli stimatori di calibrazione. Nella tabella A1, relativamente a campioni di unità elementari, si descrive il legame esistente tra alcuni degli stimatori più usati in letteratura e la classe degli stimatori di calibrazione.

Tabella A1 - Alcuni casi particolari dello stimatore di calibrazione per campioni di unità elementari

Stimatore	Gruppi di riferimento del modello	Tipo di modello		Fattore correttivo g_{ks}	Forma dello stimatore
		Valori assunti da x_k o x_{dk}	Valori assunti da c_k		
Horvitz-Thompson	No	π_k	π_k	1	\hat{Y}
Espansione per disegni semplici	No	n / N	n / N	1	$\hat{Y}_{\text{espansione}}$
Hàjek	Totale popolazione	1	1	N / \hat{N}	$\frac{\hat{Y}}{\hat{N}} N$
Rapporto semplice	Totale popolazione	x_k	x_k	X / \hat{X}	$\frac{\hat{Y}}{\hat{X}} X$
Rapporto separato	Ciascun gruppo coincide con uno strato ($d=h$)	x_{dk}	x_{dk}	X_h / \hat{X}_h	$\sum_{h=1}^H \frac{\hat{Y}_h}{\hat{X}_h} X_h$
Rapporto combinato	Totale popolazione	x_k	x_k	$X / \sum_h \hat{X}_h$	$\frac{\sum_h \hat{Y}_h}{\sum_h \hat{X}_h} X$
Rapporto combinato per sottopopolazioni	Ciascun gruppo d è costruito come aggregazione di strati	x_{dk}	x_{dk}	$X_d / \sum_{h \in d} \hat{X}_h$	$\sum_{d=1}^D \frac{\sum_{h \in d} \hat{Y}_h}{\sum_{h \in d} \hat{X}_h} X_d$
Rapporto post-stratificato*	Ciascun gruppo coincide con un post-strato ($d=a$)†	x_{dk}	x_{dk}	${}_a X / {}_a \hat{X}$	$\sum_{a=1}^A \frac{{}_a \hat{Y}}{{}_a \hat{X}} {}_a X$
Rapporto post-stratificato separato**	Ciascun gruppo coincide con una combinazione tra post-strato e strato ($d=a \cap h$)	x_{dk}	x_{dk}	${}_a X_h / {}_a \hat{X}_h$	$\sum_{a=1}^A \sum_{h=1}^H \frac{{}_a \hat{Y}_h}{{}_a \hat{X}_h} {}_a X_h$
Rapporto post-stratificato combinato**	Ciascun gruppo coincide con un post-strato ($d=a$)	x_{dk}	x_{dk}	${}_a X_h / \sum_h {}_a \hat{X}_h$	$\sum_{a=1}^A \frac{\sum_h {}_a \hat{Y}_h}{\sum_h {}_a \hat{X}_h} {}_a X$

*Utilizzato con un disegno semplice; ** utilizzato con disegno stratificato; † Il generico post-strato è indicato con a ($a=1, \dots, A$);

Gli stimatori presentati nella tabella A1 si possono agevolmente estendere ai casi di disegni a grappoli o a due o più stadi di campionamento.

A.2 Linearizzazione dello stimatore di regressione generalizzata

Per quanto illustrato nell'*appendice A.1*, una delle possibili espressioni dello stimatore di regressione generalizzata è

$$\hat{Y}_{GREG} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{B}},$$

che può essere riscritta nel seguente modo

$$\hat{Y}_{GREG} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})' \hat{\mathbf{T}}_1^{-1} \hat{\mathbf{T}}_2. \quad (\text{A.2.1})$$

La (A.2.1) evidenzia come \hat{Y}_{GREG} sia una funzione non lineare degli stimatori lineari non distorti \hat{Y} , $\hat{\mathbf{X}}$, $\hat{\mathbf{T}}_1$ e $\hat{\mathbf{T}}_2$ e rispettivamente dei totali \mathbf{Y} , \mathbf{X} , \mathbf{T}_1 e \mathbf{T}_2 .

Sia in generale $\tilde{Y} = f(\hat{\theta}_1, \dots, \hat{\theta}_q)$ uno stimatore del parametro $Y = f(\theta_1, \dots, \theta_q)$, in cui f è una funzione non lineare e il generico $\hat{\theta}_i$ è uno stimatore lineare non distorto del totale θ_i della variabile ϑ_i , ($i = 1, \dots, q$).

In presenza di funzioni non lineari, si pone il problema della determinazione della stima della media e della varianza di \tilde{Y} . Il software per il calcolo degli errori campionari risolve tale problema con il metodo della linearizzazione in serie di Taylor, il quale consiste nell'approssimare lo stimatore \tilde{Y} con una funzione lineare dei $\hat{\theta}_i$.

Per applicare il metodo è necessario che f sia differenziabile almeno fino al secondo ordine in un intorno sufficientemente ampio del punto $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)'$. Indicato $\hat{\boldsymbol{\theta}}$ con il vettore $(\hat{\theta}_1, \dots, \hat{\theta}_q)'$, lo sviluppo in serie di Taylor di \tilde{Y} intorno a $\boldsymbol{\theta}$ rispetto alle variabili $\hat{\theta}_i$ porta all'identità

$$\tilde{Y} = f(\boldsymbol{\theta}) + \sum_{i=1}^q g_i(\boldsymbol{\theta})(\hat{\theta}_i - \theta_i) + R_2, \quad (\text{A.2.2})$$

dove

$$g_i(\boldsymbol{\theta}) = \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \hat{\theta}_i} \right]_{\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}}$$

è il valore assunto dalla derivata parziale di \tilde{Y} rispetto a $\hat{\theta}_i$ calcolata nel punto $\boldsymbol{\theta}$, mentre R_2 è il resto della formula di Taylor, espresso come funzione dei termini di ordine superiore al primo. Se la dimensione campionaria n è sufficientemente elevata R_2 , può essere considerato trascurabile rispetto agli altri termini. Quindi, essendo $f(\boldsymbol{\theta}) = Y$, la (A.2.2) si può scrivere come

$$\tilde{Y} - Y \doteq \sum_{i=1}^q g_i(\boldsymbol{\theta})(\hat{\theta}_i - \theta_i). \quad (\text{A.2.3})$$

Calcolando il valore atteso in entrambi i membri, si ottiene

$$E(\tilde{Y}) - Y \doteq \sum_{i=1}^q g_i(\boldsymbol{\theta})[E(\hat{\theta}_i) - \theta_i] = 0,$$

dalla quale si deduce che \tilde{Y} è uno stimatore approssimativamente corretto di Y . Di conseguenza, elevando entrambi i membri della (A.2.3) al quadrato e passando ai valori attesi si ha

$$V(\tilde{Y}) = E(\tilde{Y} - Y)^2 \doteq V \left[\sum_{i=1}^q g_i(\boldsymbol{\theta})\hat{\theta}_i \right]. \quad (\text{A.2.4})$$

La (A.2.4) richiede il calcolo delle varianze e covarianze degli stimatori $\hat{\theta}_i$, operazione che dal punto di vista computazionale può risultare piuttosto onerosa. Per ovviare a tale inconveniente, è possibile ricorrere alla trasformata di Woodruff (1971). Infatti, l'approssimazione della varianza \tilde{Y} di data dalla (A.2.4) si può riformulare mediante la varianza dello stimatore corretto del totale

$$Z = \sum_{k \in U} z_k$$

in cui

$$z_k = \sum_{i=1}^q g_i(\boldsymbol{\theta}) \theta_{ik}$$

è il valore della trasformata di Woodruff calcolato sull'unità k , dove θ_{ik} è il valore assunto dalla variabile ϑ_i sull'unità medesima. Quindi, per la stima della varianza si utilizza l'approssimazione

$$V(\tilde{Y}) \doteq V(\hat{Z}), \quad (\text{A.2.5})$$

in cui

$$\hat{Z} = \sum_{i=1}^q g_i(\boldsymbol{\theta}) \hat{\theta}_i \quad (\text{A.2.6})$$

è uno stimatore corretto del totale Z .

Pertanto, data la variabile $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4) = (Y, \mathbf{X}, \mathbf{T}_1, \mathbf{T}_2)$, in cui \mathbf{T}_1 e \mathbf{T}_2 sono le variabili che hanno come totali rispettivamente T_1 e T_2 , applicando quanto appena visto allo stimatore di regressione generalizzata, ponendo $\boldsymbol{\theta} = (Y, \mathbf{X}, T_1, T_2)$ e $\hat{\boldsymbol{\theta}} = (\hat{Y}, \hat{\mathbf{X}}, \hat{T}_1, \hat{T}_2)$ si ha:

$$g_1(\boldsymbol{\theta}) = \left. \frac{\partial \hat{Y}_{GREG}}{\partial \hat{Y}} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 1,$$

$$g_2(\boldsymbol{\theta}) = \left. \frac{\partial \hat{Y}_{GREG}}{\partial \hat{X}_j} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = -\hat{B}_j \Big|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = -B_j, \quad j = 1, \dots, J,$$

$$g_3(\boldsymbol{\theta}) = \left. \frac{\partial \hat{Y}_{GREG}}{\partial t_{1jj'}} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = (\mathbf{X} - \hat{\mathbf{X}})' (-\hat{T}_1^{-1} \Lambda_{jj'} \hat{T}_1^{-1}) \hat{T}_2 \Big|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0, \quad j \leq j' = 1, \dots, J,$$

$$g_4(\boldsymbol{\theta}) = \left. \frac{\partial \hat{Y}_{GREG}}{\partial t_{2j}} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = (\mathbf{X} - \hat{\mathbf{X}})' \hat{T}_1^{-1} \lambda_j \Big|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} = 0, \quad j \leq j' = 1, \dots, J,$$

in cui $\Lambda_{jj'}$ è una matrice $J \times J$ con il valore 1 nella posizione (j, j') e il valore 0 altrove; λ_j è un vettore di dimensione J con il j -simo elemento pari ad 1 e tutti gli altri uguali a 0; $t_{1jj'}$ è l'elemento (j, j') della matrice \mathbf{T}_1 ; t_{2j} è l'elemento j -simo del vettore \mathbf{T}_2 .

Sostituendo le derivate $g_i(\boldsymbol{\Theta})$ ($i=1, \dots, 4$) nella (A.2.6), si ottiene

$$\hat{\mathbf{Z}} = \hat{\mathbf{Y}} - \hat{\mathbf{X}}\mathbf{B} = \sum_{k \in s} \frac{y_k - \mathbf{x}'_k \mathbf{B}}{\pi_k} = \sum_{k \in s} \frac{z_k}{\pi_k}$$

e, dunque, si è in grado di trovare l'approssimazione di $V(\hat{\mathbf{Y}}_{GREG})$ data dalla (A.2.5).

Per quanto riguarda lo stimatore della varianza di $\hat{\mathbf{Y}}_{GREG}$, una espressione generale è data da

$$var(\hat{\mathbf{Y}}_{GREG}) = var\left(\sum_{k \in s} \frac{y_k - \mathbf{x}'_k \hat{\mathbf{B}}}{\pi_k} g_{ks}\right) = var\left(\sum_{k \in s} \frac{\hat{z}_k}{\pi_k} g_{ks}\right), \quad (\text{A.2.7})$$

in cui, si introduce il termine approssimato della trasformata di Woodruff

$$\hat{z}_k = y_k - \mathbf{x}'_k \hat{\mathbf{B}} \quad (\text{A.2.8})$$

ed il fattore correttivo g_{ks} il quale permette di ottenere uno stimatore meno distorto, sotto il modello, rispetto a quello che utilizza i soli coefficienti diretti $1/\pi_k$ (Dewille e Särndal, 1992).

Si può infine notare che i valori \hat{z}_k coincidono con i termini e_k definiti nella (A.1.5). Nella precedente trattazione ci si è riferiti al caso di un modello a livello di unità elementari e di un gruppo di riferimento del modello a livello di totale popolazione. E' facile, tuttavia, adottare tale metodologia agli altri modelli descritti nella *appendice A.1*.

A.3 Lo stimatore di regressione generalizzata per i diversi disegni di campionamento

Nel presente paragrafo sono presentate le espressioni dello stimatore di regressione \hat{Y}_{GREG} , e il relativo stimatore della varianza, $var(\hat{Y}_{GREG})$, nei diversi disegni di campionamento con e senza reimmissione. Per non appesantire eccessivamente tale trattazione si esaminano direttamente le strategie campionarie che adottano un disegno stratificato, tralasciando l'analisi del caso in cui la popolazione non sia suddivisa in strati. Quest'ultimo caso, tuttavia, è facilmente riconducibile al campionamento stratificato considerando una popolazione costituita da un unico strato.

A.3.1 Campionamento di unità elementari con probabilità d'inclusione costanti

Sia U una popolazione suddivisa in H strati e si indichi con:

- h ($h=1, \dots, H$) l'indice del generico strato costituito da N_h unità, dove $\sum_h N_h = N$;
- k ($k=1, \dots, N_h$) l'indice della generica unità finale di campionamento appartenente allo strato h ;

Il parametro da stimare si può in questo caso esprimere come

$$Y = \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} ,$$

dove y_{hk} rappresenta il valore assunto dalla variabile Y sull'unità elementare k inclusa nello strato h .

Si supponga di aver estratto da U , attraverso un disegno casuale stratificato, un campione s , in cui per ciascuno strato h la selezione delle n_h unità ($\sum_h n_h = n$) sia stata effettuata con reimmissione e probabilità uguali. In tale contesto lo stimatore di regressione generalizzata per il totale Y si può scrivere come

$$\hat{Y}_{GREG} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_{hk} g_{hk} \quad (A.3.1)$$

in cui il termine N_h/n_h rappresenta il coefficiente diretto dell'unità k appartenente allo strato h e g_{hk} è un fattore correttivo ottenuto mediante l'espressione (A.1.12) o alternativamente dalla (A.1.14), a seconda del tipo di gruppo di riferimento del modello adottato.

In base alla (A.2.7), il software calcola la stima della varianza dello stimatore \hat{Y}_{GREG} mediante l'espressione

$$\text{var}(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(\frac{\hat{z}_{hk} g_{hk}}{\pi_{hk}} - \bar{\bar{Z}}_h \right)^2 = \sum_{h=1}^H \frac{N_h^2}{n_h} \frac{1}{n_h - 1} \sum_{k=1}^{n_h} \left(\hat{z}_{hk} g_{hk} - \bar{\bar{Z}}_h \right)^2 \quad (A.3.2)$$

in cui \hat{z}_{hk} è la trasformata di y_{hk} data dall'espressione (A.2.8) e dove

$$\bar{\bar{Z}}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} \hat{z}_{hk} g_{hk} .$$

Se la selezione delle unità nel campione avviene senza reimmissione, lo stimatore del parametro Y è dato sempre dalla (A.3.1), mentre la stima della varianza è calcolata tramite l'espressione

$$\begin{aligned} \text{var}(\hat{Y}_{GREG}) &= \sum_{h=1}^H \frac{N_h^2}{n_h} \frac{1}{n_h - 1} \left(1 - \frac{n_h}{N_h} \right) \sum_{k=1}^{n_h} \left(\hat{z}_{hk} g_{hk} - \bar{\bar{Z}}_h \right)^2 = \\ &= \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} \frac{1}{n_h - 1} \sum_{k=1}^{n_h} \left(\hat{z}_{hk} g_{hk} - \bar{\bar{Z}}_h \right)^2 \end{aligned} \quad (A.3.3)$$

E' da sottolineare che nella (A.3.3) compare esplicitamente il termine N_h , a differenza di quanto avviene per la (A.3.2) in cui non è richiesta la conoscenza diretta di N_h in quanto sostituendo π_{hk} con n_h/N_h si ottiene la prima delle (A.3.2) che non dipende da N_h .

In base a tale considerazione per calcolare la (A.3.3) bisogna, dunque, conoscere la numerosità dello strato. Tuttavia, nella progettazione del

software si è deciso di non richiedere questa ulteriore informazione all'utente e la formula (A.3.3) è calcolata sostituendo N_h con la \hat{N}_h stima ottenuta con i pesi diretti. Tale stima riporta esattamente al totale N_h quando tutte le unità del campione hanno risposto. In presenza del fenomeno della mancata risposta totale, nel caso in cui sono stati utilizzati come coefficienti iniziali di input i coefficienti diretti senza la correzione per mancata risposta totale, la quantità \hat{N}_h sottostima il totale N_h . In presenza di mancata risposta totale, si consiglia pertanto di utilizzare nel software i coefficienti diretti corretti per mancata risposta totale.

La (A.3.2) e la (A.3.3) rappresentano una stima corretta della varianza se \hat{Y}_{GREG} è uno stimatore lineare, mentre sono consistenti per il disegno (*design consistent*) e sono approssimativamente corretti rispetto al modello di regressione sottostante se lo stimatore \hat{Y}_{GREG} non è lineare (Särndal et al., 1992 pag.238; Särndal et al., 1989).

A.3.2 Campionamento a grappoli con probabilità d'inclusione costanti

Si definisca con U l'universo di riferimento dei grappoli (già introdotto nel paragrafo A.1.2) con U_I suddiviso in H strati e in relazione al generico strato h si indichi con:

- i ($i=1, \dots, N_h$) l'indice della generico grappolo di unità elementari;
- k ($k=1, \dots, M_{hi}$) l'indice della generica unità elementare appartenente al grappolo i dello strato h .

Inoltre, si denoti sinteticamente con (hik) la generica unità elementare k inclusa nel grappolo i dello strato h .

In questo caso il parametro si può rappresentare come

$$Y = \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik} ,$$

dove y_{hik} è il valore della variabile Y osservato sull'unità elementare (hik) .

Sia s un campione di n grappoli ottenuto attraverso un disegno casuale stratificato, in cui per ciascuno strato si estraggono con reimmissione e

probabilità uguali n_b grappoli. In questo tipo di disegno, che prevede un solo stadio di selezione ed in cui si selezionano grappoli di unità elementari, le unità primarie di campionamento coincidono con le unità finali di campionamento che sono rappresentate dai grappoli di unità elementari.

Nel campionamento a grappoli la definizione dello stimatore di regressione generalizzata varia a seconda del livello del modello utilizzato. La scelta del livello, influisce sulla forma dello stimatore nella definizione del fattore correttivo. In generale lo stimatore è espresso come

$$\hat{Y}_{GREG} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{M_{hi}} y_{hik} g_{hik} , \quad (\text{A.3.4})$$

in cui per il modello a livello di unità elementari, g_{hik} è dato dalla:

- (1) (A.1.12), se si utilizza un unico gruppo di riferimento del modello, che coincide con l'intera popolazione;
- (2) (A.1.14), se si utilizzano D ($d=1, \dots, D$) gruppi di riferimento del modello.

Per il modello a livello di cluster si ha che g_{hik} è dato dalla

- (3) (A.1.17), se si utilizza un unico gruppo di riferimento del modello, che coincide con l'intera popolazione;
- (4) (A.1.18), se si utilizzano D ($d=1, \dots, D$) gruppi di riferimento del modello.

Adattando la (A.2.7) a questo disegno di campionamento, la stima della varianza dello stimatore, \hat{Y}_{GREG} definito dalla (A.3.4), è calcolata dal software con la formula seguente

$$var(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{N_h^2}{n_h} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\tilde{Z}_{hi} - \tilde{\bar{Z}}_h)^2 , \quad (\text{A.3.5})$$

essendo

$$\tilde{Z}_{hi} = \sum_{k=1}^{M_{hi}} \hat{z}_{hik} g_{hik} , \quad \tilde{\bar{Z}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{Z}_{hi} .$$

Se la selezione dei grappoli avviene senza reimmissione, lo stimatore è

sempre espresso dalla (A.3.4), mentre la stima della sua varianza si ottiene con

$$\text{var}(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{n_h} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\tilde{Z}_{hi} - \tilde{\bar{Z}}_h \right)^2 . \quad (\text{A.3.6})$$

Nella (A.3.6) valgono le stesse considerazioni esposte in relazione alla (A.3.3) per quanto riguarda il termine N_h .

Le espressioni (A.3.5) e (A.3.6) rappresentano stimatori corretti (o approssimativamente corretti se la funzione è non lineare) della varianza campionaria, nel caso in cui si adotta uno stimatore \hat{Y}_{GREG} espresso dalla (A.3.4).

A.3.3 Campionamento di unità elementari con probabilità d'inclusione variabili

In presenza di un disegno con probabilità di inclusione variabili lo stimatore del totale Y si presenta come:

$$\hat{Y}_{GREG} = \sum_{h=1}^H \sum_{k=1}^{n_h} \frac{y_{hk} g_{hk}}{\pi_{hk}} , \quad (\text{A.3.7})$$

in cui si è indicato con π_{hk} la probabilità d'inclusione dell'unità k nello strato h e g_{hk} il fattore correttivo ottenuto tramite la (A.1.12) o la (A.1.14). La (A.3.7) rappresenta un'espressione più generale della (A.3.1) ed è valida per un disegno di campionamento con o senza reimmissione.

Per quanto riguarda la stima della varianza di \hat{Y}_{GREG} il software non opera distinzioni tra selezione del campione con reimmissione e senza reimmissione come avviene, invece, in presenza di un disegno con probabilità di inclusione costanti, all'interno degli strati.

Secondo la (A.2.7), lo stimatore adottato è

$$\text{var}(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{k=1}^{n_h} \left(\frac{\hat{z}_{hk} g_{hk}}{\pi_{hk}} - \tilde{\bar{Z}}_h \right)^2 , \quad (\text{A.3.8})$$

essendo

$$\tilde{\bar{Z}}_h = \frac{1}{n_h} \sum_{k=1}^{n_h} \frac{\hat{z}_{hk} g_{hk}}{\pi_{hk}} .$$

Lo stimatore (A.3.8) risulta corretto (o approssimativamente corretto se \hat{Y}_{GREG} è non lineare) nel caso in cui il campione sia stato selezionato con reimmissione, mentre risulta distorto se il campione è stato selezionato senza reimmissione, determinando delle stime approssimate per eccesso. Tuttavia, è necessario sottolineare che la distorsione è trascurabile quando il tasso di campionamento all'interno degli strati è “piccolo” (Wolter, 1985).

La scelta di non utilizzare lo stimatore corretto (o approssimativamente corretto se \hat{Y}_{GREG} è non lineare) della varianza quando la selezione delle unità è senza reimmissione, è dettata dalla difficoltà di calcolo delle probabilità di inclusione di secondo ordine delle unità, le quali sono necessarie per definire tale stimatore. Ulteriori considerazioni sull'uso dell'espressione (A.3.8) per disegni senza reimmissione sono state evidenziate nel capitolo 4.

A.3.4 Campionamento a grappoli con probabilità d'inclusione variabili

Lo stimatore \hat{Y}_{GREG} in tale contesto assume la forma:

$$\hat{Y}_{GREG} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{k=1}^{M_{hi}} \frac{y_{hik} g_{hik}}{\pi_{hik}}, \quad (\text{A.3.9})$$

essendo per l'unità (hik) :

- π_{hik} la probabilità d'inclusione costante per tutte le unità elementari appartenenti al grappolo i dello strato h , e pari alla probabilità di inclusione π_{hi} dello stesso grappolo i ;
- g_{hik} , il fattore correttivo che si può esprimere alternativamente con la (A.1.12), la (A.1.14), la (A.1.17) o la (A.1.18) a seconda che si usino o no i gruppi di riferimento ed a seconda del livello del modello prescelto.

Per gli analoghi motivi descritti nel caso del campionamento di unità elementari con probabilità d'inclusione variabili, il software applica la stima della varianza del caso con reimmissione, anche quando si è adottato uno schema di selezione senza reimmissione; per la stima della varianza dello stimatore (A.3.9), la formula impiegata è

$$\text{var}(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\tilde{Z}_{hi} - \tilde{Z}_h)^2, \quad (\text{A.3.10})$$

essendo

$$\tilde{Z}_{hi} = \sum_{k=1}^{M_{hi}} \frac{1}{\pi_{hik}} \hat{z}_{hik} g_{hik}, \quad \tilde{Z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{Z}_{hi}.$$

La (A.3.10) risulta corretta quando la selezione avviene con reimmissione ed è distorta quando la selezione dei grappoli è senza reimmissione.

A.3.5 Campionamento a due o più stadi

Il software per il calcolo degli errori di campionamento è progettato, principalmente, per la stima della varianza dello stimatore di regressione generalizzata per un disegno di campionamento a due o più stadi, con probabilità di selezione variabile delle unità di primo stadio (UPS). Ciò in quanto tra i disegni a due o più stadi, è quello maggiormente utilizzato nelle indagini effettive su larga scala.

Si consideri in una prima fase un disegno a due stadi, e sia, quindi, U l'universo di riferimento delle UPS suddiviso in H strati e in relazione al generico strato h si indichi con:

- i ($i=1, \dots, N_h$) l'indice della generica UPS;
- k ($k=1, \dots, M_{hi}$) l'indice della generica unità elementare di secondo stadio (USS) appartenente all'unità primaria i .

Inoltre, analogamente a quanto visto nel precedente paragrafo, si denoti sinteticamente con (hik) la generica USS k inclusa nella UPS i dello strato h .

Il parametro da stimare è, quindi, dato da

$$Y = \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik},$$

dove y_{hik} è il valore della variabile Y osservato sull'unità elementare (hik) .

Prendiamo in esame il caso della selezione delle UPS con probabilità variabili e siano rispettivamente n_h il numero di UPS selezionate nello

strato h e m_{hi} il numero delle USS selezionate nella UPS i dello strato h .

In tale contesto lo stimatore \hat{Y}_{GREG} , sia nel caso di selezione della UPS con reimmissione che in quello senza reimmissione, è dato dalla seguente espressione

$$\hat{Y}_{GREG} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} \frac{y_{hik} g_{hik}}{\pi_{hik}} \quad (A.3.11)$$

dove la probabilità di inclusione π_{hik} della generica USS (hik) è data dal prodotto tra la probabilità di inclusione π_{hi} della UPS (hi) e la probabilità di inclusione condizionata $\pi_{k|hi}$ della stessa USS (hik), dato che al primo stadio è stata selezionata la UPS (hi).

La stima della varianza di \hat{Y}_{GREG} calcolata dal software con la stessa formula, sia per la selezione con reimmissione che per quella senza reimmissione delle UPS, è data da

$$var(\hat{Y}_{GREG}) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\tilde{Z}_{hi} - \bar{\tilde{Z}}_h \right)^2, \quad (A.3.12)$$

essendo

$$\tilde{Z}_{hi} = \sum_{k=1}^{m_{hi}} \frac{1}{\pi_{hik}} \hat{z}_{hik} g_{hik}, \quad \bar{\tilde{Z}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{Z}_{hi}.$$

In un disegno che prevede per le UPS probabilità di inclusione di primo ordine variabili, la (A.3.12) rappresenta uno stimatore corretto (o approssimativamente corretto se \hat{Y}_{GREG} è non lineare) nel caso che il campione sia stato selezionato con reimmissione e presenta, invece, una distorsione positiva qualora la selezione delle UPS sia stata compiuta senza reimmissione.

In quest'ultimo caso l'uso della (A.3.12) è giustificato dalla difficoltà di calcolo delle probabilità di inclusione di secondo ordine delle UPS, richieste per definire lo stimatore corretto della varianza (o approssimativamente corretto se \hat{Y}_{GREG} è non lineare).

Nel caso in cui le UPS siano estratte con probabilità di inclusione costanti, il software utilizza sempre la (A.3.12) che è uno stimatore corretto per la selezione delle UPS con reimmissione e distorto positivamente per la selezione delle UPS senza reimmissione.

Per disegni a tre o più stadi di campionamento non si presentano differenze sostanziali. Gli stadi di campionamento ulteriori al secondo sono integrati nella (A.3.11) attraverso l'inserimento di altre sommatorie per tenere conto delle unità selezionate nel campione negli stadi successivi, mentre la stima della varianza si ottiene sempre con la (A.3.12).

A.4 La costruzione dei data-set di input per definire i gruppi di riferimento

Nella presente appendice sono trattati con maggiore approfondimento i criteri di costruzione del *data-set* di input ed, in particolare, le alternative possibili per definire le variabili POP_PIAN e le variabili X_j ($j=1, \dots, J$) in relazione al processo di stima per calcolare i coefficienti finali di output. Tali criteri sono stati introdotti nel *paragrafo 1.3*.

A.4.1 Costruzione dei gruppi di riferimento: caso I

Gruppi di riferimento definiti su sottopopolazioni pianificate ottenute marginalizzando alcune variabili che contribuiscono a definire la stratificazione e con variabili ausiliarie X quantitative o qualitative dicotomiche

Questo primo approfondimento sulla costruzione del *data-set* di input prevede due ipotesi di base:

- la prima richiede che la variabile di stratificazione sia multivariata¹¹ e che le sottopopolazioni pianificate, definite come aggregazioni di strati, siano il risultato di un processo di aggregazione rispetto ad una o più variabili che identificano gli strati stessi;

¹¹ Per variabile multivariata si intende che ciascuna modalità può essere definita come la combinazione delle modalità di due o più variabili.

- la seconda ipotesi suppone che le variabili qualitative interessate dal processo di calibrazione siano dicotomiche del tipo presenza/assenza, sì/no, 0/1, ecc..

Nella prima ipotesi rientrano anche le strategie di campionamento in cui le sottopopolazioni pianificate coincidono con gli strati. In tal caso non è necessario distinguere le variabili di stratificazione tra semplici e multivariate.

Per chiarire quali condizioni sono richieste nelle due ipotesi si consideri l'esempio seguente.

Esempio A.4.1:

Sia dato un campione d'individui stratificato sulla base di una variabile che è ottenuta dalla combinazione delle quattro variabili descritte nella tabella A.4.1.

Tabella A.4.1 - Variabili che descrivono la stratificazione

Variabili che definiscono la stratificazione	Simbolo variabile	Numero di modalità	Simbolo numero di modalità	Modalità
Sesso	s_1	2	S_1	uomo; donna
Classe di età	s_2	4	S_2	0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre
Stato civile	s_3	2	S_3	sposato; non sposato
Ripartizione geografica	s_4	3	S_4	nord; centro; sud

Si considerino anche tre variabili di post-stratificazione (che non rientrano nella definizione degli strati del disegno), descritte nella tabella A.4.2, per le quali sono noti i totali di alcune variabili ausiliarie utilizzate per definire lo stimatore di ponderazione vincolata adottato.

Tabella A.4.2 - Variabili dell'esempio che definiscono sottopopolazioni non pianificate (variabili di post-stratificazione)

Variabili di post-stratificazione	Simbolo	Numero di modalità	Simbolo numero di modalità	Modalità
Settore di attività economica in cui lavora l'individuo	v_1	3	Q_1	agricoltura; industria; terziario;
Professione	v_2	4	Q_2	operaio; impiegato; dirigente; altro
Titolo di studio	v_3	4	Q_3	licenza elementare; licenza media; diploma di scuola superiore; laurea universitaria

Si abbiano inoltre quattro variabili ausiliarie, definite nella tabella A.4.3, utilizzate nello stimatore di ponderazione vincolata, per le quali si conoscono i totali di popolazione su alcune partizioni in gruppi di riferimento della popolazione.

Tabella A.4.3 - Variabili dell'esempio che presentano dei totali noti a livello di sottopopolazioni

Variabili ausiliarie	Simbolo	Numero di modalità	Modalità
Indicatore di presenza dell'unità nella sottopopolazione.	x_1	2	appartenente alla sottopopolazione (1), non appartenente alla sottopopolazione (0). Per come viene definito il data-set la variabile è sempre pari a "1".
Indicatore di proprietà dell'abitazione	x_2	2	proprietario (1), non proprietario (0)
Numero di figli	x_3	-	-
Reddito individuale	x_4	-	-

Infine, si considerino cinque differenti partizioni della popolazione in gruppi di riferimento, descritte nella tabella A.4.4, in cui sono noti i totali di popolazione per alcune delle variabili ausiliarie introdotte nella tabella A.4.3.

Tabella A.4.4 – Descrizione delle partizioni in gruppi di riferimento prese in considerazione nell'esempio

Partizioni	Simbolo	Variabili che definiscono i gruppi di riferimento della partizione	Numero dei gruppi di riferimento nella partizione	Variabili ausiliarie per le quali si hanno i totali noti
Prima partizione	P_1	s_1, s_2, s_3, s_4	D_1	x_1
Seconda partizione	P_2	s_1, s_2, v_2	D_2	x_2, x_4
Terza partizione	P_3	s_1, s_2, s_4, v_3	D_3	x_3
Quarta partizione	P_4	s_1, s_2, v_1	D_4	x_3, x_4
Quinta partizione	P_5	s_1, s_2, s_4, v_1	D_5	x_4

Per rendere chiaro quali sono le informazioni contenute nella tabella A.4.4 si osservi, ad esempio, la prima riga relativa alla prima partizione in gruppi di riferimento. Ciascun gruppo di riferimento di questa partizione è identificato da una particolare combinazione delle modalità di tutte le variabili che definiscono gli strati del disegno. Per i gruppi di questa prima partizione il totale utilizzato a livello di stimatore di ponderazione vincolata è il totale della popolazione.

La seconda partizione (seconda riga) è costituita dai gruppi di riferimento identificati dall'incrocio delle modalità della variabile sesso, classe di età e della professione. In tali gruppi sono noti il totale di sottopopolazione degli individui possessori di una abitazione e il totale di sottopopolazione dei redditi individuali.

Per concludere questa breve descrizione delle caratteristiche delle cinque partizioni si può osservare che: la prima partizione presenta come sottopopolazioni pianificate i singoli strati; le restanti partizioni si basano, invece, su sottopopolazioni pianificate ricavate marginalizzando su una o più variabili che definiscono la stratificazione. In particolare, nella seconda partizione si marginalizza sulle variabili s_3 , s_4 , nella terza partizione si marginalizza sulla variabile s_3 , e così via nelle altre due partizioni.

L'utente per indicare al software quali sono le partizioni in gruppi di riferimento della popolazione obiettivo utilizzate dallo stimatore di ponderazione vincolata, deve agire sulla definizione delle modalità della variabile POP_PLAN, sulla costruzione di un certo numero di variabili X_j (si veda il paragrafo 1.3) e sulla definizione dei valori che possono assumere queste ultime. A tale scopo si può adottare una delle tre alternative introdotte nel paragrafo 1.3. Nelle tabelle A.4.5, A.4.6 e A.4.7 è descritto come costruire il data set "dati campionari".

Tabella A.4.5 – Costruzione del data-set “dati campionari” secondo lo schema A

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili X_j
Numero Modalità	POP_PIAN	1	1	Nessuna variabile identifica POP_PIAN
		↓	↓	↓
Numero variabili	X_j Per tenere conto dei totali della variabile x_1 sulla partizione P_1	48 X_1, \dots, X_{48}	\times $S_1 \times S_2 \times S_3 \times S_4$	s_1, s_2, s_3, s_4
Numero variabili	X_j Per tenere conto dei totali della variabile x_2 sulla partizione P_2	32 X_{49}, \dots, X_{80}	$S_1 \times S_2 \times Q_2$	s_1, s_2, v_2
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_3	96 X_{81}, \dots, X_{176}	$S_1 \times S_2 \times S_4 \times Q_3$	s_1, s_2, s_4, v_3
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_4	24 X_{177}, \dots, X_{200}	$S_1 \times S_2 \times Q_1$	s_1, s_2, v_1
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_2	32 X_{201}, \dots, X_{232}	$S_1 \times S_2 \times Q_2$	s_1, s_2, v_2
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_4	24 X_{233}, \dots, X_{256}	$S_1 \times S_2 \times Q_1$	s_1, s_2, v_1
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_5	72 X_{257}, \dots, X_{328}	$S_1 \times S_2 \times S_4 \times Q_1$	s_1, s_2, s_4, v_1
Numero totale di variabili X_j nel data-set di input		328 X_1, \dots, X_{328}		

Tabella A.4.6 – Costruzione del data-set “dati campionari” secondo lo schema B

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili X_j
Numero Modalità	POP_PIAN	8	$S_1 \times S_2$	s_1, s_2
		↓	↓	↓
Numero variabili	X_j Per tenere conto dei totali della variabile x_1 sulla partizione P_1	6 X_1, \dots, X_6	$S_3 \times S_4$	s_3, s_4
Numero variabili	X_j Per tenere conto dei totali della variabile x_2 sulla partizione P_2	4 X_7, \dots, X_{10}	Q_2	v_2
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_3	12 X_{11}, \dots, X_{22}	$S_4 \times Q_3$	s_4, v_3
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_4	3 X_{23}, \dots, X_{25}	Q_1	v_1
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_2	4 X_{26}, \dots, X_{29}	Q_2	v_2
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_4	3 X_{30}, \dots, X_{32}	Q_1	v_1
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_5	9 X_{33}, \dots, X_{41}	$S_4 \times Q_1$	s_4, v_1
Numero totale di variabili X_j nel data-set di input		41 X_1, \dots, X_{41}		

Tabella A.4.7 – Costruzione del data-set “dati campionari” secondo lo schema C

	Variabili di input	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j	Numero delle modalità della variabile POP_PIAN e numero delle variabili X_j (simboli)	Variabili del disegno che identificano la variabile POP_PIAN e le variabili X_j
Numero Modalità	POP_PIAN	2 (oppure 4)	S_1 (oppure S_2)	s_1 (oppure s_2)
Numero variabili	X_j Per tenere conto dei totali della variabile x_1 sulla partizione P_1	\Downarrow 24 X_1, \dots, X_{24} (oppure 12 X_1, \dots, X_{12})	\Downarrow $S_2 \times S_3 \times S_4$ (oppure $S_1 \times S_3 \times S_4$)	\Downarrow s_2, s_3, s_4 (oppure s_1, s_3, s_4)
Numero variabili	X_j Per tenere conto dei totali della variabile x_2 sulla partizione P_2	16 X_{25}, \dots, X_{40} (oppure 8 X_{13}, \dots, X_{20})	$S_2 \times Q_2$ (oppure $S_1 \times Q_2$)	s_2, v_2 (oppure s_1, v_2)
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_3	48 X_{41}, \dots, X_{88} (oppure 24 X_{21}, \dots, X_{44})	$S_2 \times S_4 \times Q_3$ (oppure $S_1 \times S_4 \times Q_3$)	s_2, s_4, v_3 (oppure s_1, s_4, v_3)
Numero variabili	X_j Per tenere conto dei totali della variabile x_3 sulla partizione P_4	12 X_{89}, \dots, X_{100} (oppure 6 X_{45}, \dots, X_{50})	$S_2 \times Q_1$ (oppure $S_1 \times Q_1$)	s_2, v_1 (oppure s_1, v_1)
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_2	16 X_{101}, \dots, X_{116} (oppure 8 X_{51}, \dots, X_{58})	$S_2 \times Q_2$ (oppure $S_1 \times Q_2$)	s_2, v_2 (oppure s_1, v_2)
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_4	12 X_{117}, \dots, X_{128} (oppure 6 X_{59}, \dots, X_{64})	$S_2 \times Q_1$ (oppure $S_1 \times Q_1$)	s_2, v_1 (oppure s_1, v_1)
Numero variabili	X_j Per tenere conto dei totali della variabile x_4 sulla partizione P_5	36 X_{129}, \dots, X_{164} (oppure 18 X_{65}, \dots, X_{82})	$S_2 \times S_4 \times Q_1$ (oppure $S_1 \times S_4 \times Q_1$)	s_2, s_4, v_1 (oppure s_1, s_4, v_1)
Numero totale di variabili X_j nel data-set di input		164 X_1, \dots, X_{164} (oppure 82 X_1, \dots, X_{82})		

Le informazioni contenute nelle tre tabelle sono le seguenti:

□ *Schema A (tabella A.4.5);*

- 1° la variabile POP_PLAN ha una sola modalità. Tutti i record presentano un valore costante della variabile;*
- 2° sono presenti le variabili X1, ..., X328. L'insieme di queste variabili è suddiviso in sette sottoinsiemi:*
 - 3° sottoinsieme che raggruppa le variabili X1, ..., X48: queste variabili identificano i valori della variabile x1 sulla partizione P1; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s1, s2, s3, s4 che si presenta sul record corrispondente;*
 - 4° sottoinsieme che raggruppa le variabili X49, ..., X80: queste variabili identificano i valori della variabile X2 sulla partizione P2; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili s1, s2, v2 che si presenta sul record corrispondente;*
 - 5° sottoinsieme che raggruppa le variabili X81, ..., X176: queste variabili identificano i valori della variabile X3 sulla partizione P3; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s1, s2, s4, v3 che si presenta sul record corrispondente;*
 - 6° sottoinsieme che raggruppa le variabili X177, ..., X200: queste variabili identificano i valori della variabile X3 sulla partizione P4; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s1, s2, v1 che si presenta sul record corrispondente;*
 - 7° sottoinsieme che raggruppa le variabili X201, ..., X232: queste variabili identificano i valori della variabile X4 sulla partizione P2; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identifica-*

to dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_1 , s_2 , v_2 che si presenta sul record corrispondente;

- 8° sottoinsieme che raggruppa le variabili X_{233} , ..., X_{256} : queste variabili identificano i valori della variabile X_4 sulla partizione P_4 ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_1 , s_2 , v_1 , che si presenta sul record corrispondente;*
- 9° sottoinsieme che raggruppa le variabili X_{257} , ..., X_{328} : queste variabili identificano i valori della variabile X_4 sulla partizione P_5 ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_1 , s_2 , s_4 , v_1 che si presenta sul record corrispondente;*

□ *Schema B (tabella A.4.6);*

- 1° le modalità assunte dalla variabile POP_PLAN identificano le differenti combinazioni delle modalità delle variabili s_1 , s_2 . In particolare, ciascun record presenta sulla variabile POP_PLAN la modalità che identifica la combinazione di s_1 , s_2 presente nel record stesso.*
- 2° sono presenti le variabili X_1 , ..., X_{41} . L'insieme di queste variabili è suddiviso in sette sottoinsiemi:*
- 3° sottoinsieme che raggruppa le variabili X_1 , ..., X_6 : queste variabili identificano i valori della variabile X_1 sulla partizione P_1 ; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_3 , s_4 che si presenta sul record corrispondente;*
- 4° sottoinsieme che raggruppa le variabili X_7 , ..., X_{10} : queste variabili identificano i valori della variabile X_2 sulla partizione P_2 ; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili v_2 che si presenta sul record corrispondente;*

- 5° sottoinsieme che raggruppa le variabili $X11, \dots, X22$; queste variabili identificano i valori della variabile $X3$ sulla partizione $P3$; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili $s4, v3$ che si presenta sul record corrispondente;
- 6° sottoinsieme che raggruppa le variabili $X23, \dots, X25$: queste variabili identificano i valori della variabile $X3$ sulla partizione $P4$; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili $v1$ che si presenta sul record corrispondente;
- 7° sottoinsieme che raggruppa le variabili $X26, \dots, X29$: queste variabili identificano i valori della variabile $X4$ sulla partizione $P2$; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili $v2$ che si presenta sul record corrispondente;
- 8° sottoinsieme che raggruppa le variabili $X30, \dots, X32$: queste variabili identificano i valori della variabile $X4$ sulla partizione $P4$; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili $v1$ che si presenta sul record corrispondente;
- 9° sottoinsieme che raggruppa le variabili $X33, \dots, X41$: queste variabili identificano i valori della variabile $X4$ sulla partizione $P5$; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili $s4, v1$ che si presenta sul record corrispondente.

□ Schema C (tabella A.4.7);

Relativamente allo schema C la tabella rileva l'esistenza di due possibili alternative. La prima definisce le modalità della variabile POP_PLAN in base alla variabile $s1$, la seconda, invece, sulla variabile $s2$. Descrivendo la prima delle due alternative si ha che:

- 1° *le modalità assunte dalla variabile POP_PLAN identificano (possono anche coincidere) le modalità delle variabili s1. In particolare, ciascun record presenta sulla variabile POP_PLAN la modalità che identifica la modalità di s1 che si presenta nel record stesso.*
- 2° *sono presenti le variabili X1, ..., X164. L'insieme di queste variabili è suddiviso in sette sottoinsiemi:*
- 3° *sottoinsieme che raggruppa le variabili X1, ..., X24: queste variabili identificano i valori della variabile X1 sulla partizione P1; in particolare per ciascun record una sola di queste variabili è pari a "1" e le altre sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s2, s3, s4 che si presenta sul record corrispondente;*
- 4° *sottoinsieme che raggruppa le variabili X25, ..., X40: queste variabili identificano i valori della variabile X2 sulla partizione P2; in particolare per ciascun record una sola di queste variabili può essere pari a "1" e ciò accade quando il record è relativo ad un individuo che possiede un'abitazione, mentre le altre sono nulle. La variabile che può essere pari a "1" è quella identificata dalla combinazione delle modalità delle variabili s2, v2 che si presenta sul record corrispondente;*
- 5° *sottoinsieme che raggruppa le variabili X41, ..., X88; queste variabili identificano i valori della variabile X3 sulla partizione P3; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s2, s4, v3 che si presenta sul record corrispondente;*
- 6° *sottoinsieme che raggruppa le variabili X89, ..., X100: queste variabili identificano i valori della variabile X3 sulla partizione P4; in particolare per ciascun record una sola di queste variabili è pari al numero dei figli che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s2, v1, che si presenta sul record corrispondente;*
- 7° *sottoinsieme che raggruppa le variabili X101, ..., X116: queste variabili identificano i valori della variabile X4 sulla partizione P2; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s2, v2 che si presenta sul record corrispondente;*

- 8° sottoinsieme che raggruppa le variabili X_{117}, \dots, X_{128} : queste variabili identificano i valori della variabile X_4 sulla partizione P_4 ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_2, v_1 che si presenta sul record corrispondente;
- 9° sottoinsieme che raggruppa le variabili X_{129}, \dots, X_{164} : queste variabili identificano i valori della variabile X_4 sulla partizione P_5 ; in particolare per ciascun record una sola di queste variabili è pari al reddito che ha l'individuo identificato dal record stesso, mentre le altre variabili sono nulle. La variabile non nulla è quella identificata dalla combinazione delle modalità delle variabili s_2, s_4, v_1 che si presenta sul record corrispondente.

Per rendere più generale la descrizione vista nell'esempio A.4.1 dei tre schemi di costruzione di un *data-set* di input, è necessario definire una simbologia, in parte già introdotta nell'esempio stesso, per identificare le variabili che rappresentano gli strati (tabella A.4.8), i post-strati (tabella A.4.9) e le variabili di cui si usano i totali noti a livello di stimatore (tabella A.4.10). Relativamente a queste ultime, si considerano, per il momento, le variabili quantitative e le variabili qualitative dicotomiche del tipo presenza/assenza, sì/no, 0/1.

Tabella A.4.8 – Definizione simbolica delle variabili che identificano uno strato

Variabile	s_1	...	s_a	...	s_A
Numero di modalità	S_1	...	S_a	...	S_A

Tabella A.4.9 – Definizione simbolica delle variabili di post-stratificazione

Variabile	V_1	...	v_b	...	v_B
Numero di modalità	Q_1	...	Q_b	...	Q_B

Tabella A.4.10 – Definizione simbolica delle variabili ausiliarie di cui si utilizzano i totali noti a livello di stimatore

Variabile	x_1	...	x_t	...	x_T
-----------	-------	-----	-------	-----	-------

In base alla notazione presentata nelle tabelle A.4.8 e A.4.9, una generica partizione P_i ($i=1, \dots, I$), è definibile da un sottoinsieme \underline{s}^i composto da alcune delle variabili s_a ($a=1, \dots, A$) e da una variabile di post-stratificazione v^i che coincide con una delle variabili v_b ($b=1, \dots, B$). Tale partizione, come è illustrato nella tabella A.4.11, è composta da $\underline{S}^i \times Q^i$ gruppi di riferimento, dove \underline{S}^i è il numero di combinazioni di modalità delle variabili contenute nel sottoinsieme \underline{s}^i , mentre Q^i è il numero di modalità di v^i .

Tabella A.4.11 – Descrizione simbolica delle partizioni in gruppi di riferimento di una popolazione oggetto d'indagine

Indicatore di partizione	P_1	...	P_i	...	P_I
Insieme di variabili di stratificazione che identificano la partizione	\underline{s}^1	...	\underline{s}^i	...	\underline{s}^I
Numero delle combinazioni di modalità delle variabili di stratificazione che identificano la partizione	\underline{S}^1	...	\underline{S}^i	...	\underline{S}^I
Variabile di post-stratificazione che identifica la partizione	v^1	...	v^i	...	v^I
Numero delle modalità della variabile di post-stratificazione che identifica la partizione	Q^1	...	Q^i	...	Q^I
Numero dei gruppi di riferimento della partizione	$\underline{S}^1 \times Q^1$...	$\underline{S}^i \times Q^i$...	$\underline{S}^I \times Q^I$

Dati questi elementi, si indichi con \underline{s} il sottoinsieme delle variabili di stratificazione che sono contenute in tutti gli insiemi \underline{s}^i . Inoltre sia \underline{S} il numero di combinazioni delle modalità delle variabili in \underline{s} . Pertanto, per la generica partizione P_i , il numero dei gruppi di riferimento si può denotare con il prodotto $\underline{S} \times \bar{\underline{S}}^i \times Q^i$ in cui $\bar{\underline{S}}^i$ è il numero delle combinazioni delle modalità dell'insieme di variabili $\bar{\underline{S}}^i$ incluse in \underline{s}^i ed escluse da \underline{s} , avendo, quindi, $\underline{s}^i = \underline{s} \cup \bar{\underline{S}}^i$.

Considerata la simbologia sopra introdotta, è possibile, allora, dare una struttura generale per definire lo schema A e lo schema B (si veda tabella A.4.12).

Per impostare il *data-set* di input secondo lo schema C è necessario definire con ${}^c\underline{s}$ e $\bar{{}^c\underline{s}}$ due sottoinsiemi di variabili tra loro disgiunti la cui unio-

ne riporta ad \underline{s} . Si indichi con ${}^c\underline{S}$ il numero delle combinazioni delle modalità delle variabili in ${}^c\underline{s}$ e con ${}^{\bar{c}}\underline{S}$ il numero delle combinazioni delle modalità delle variabili in ${}^{\bar{c}}\underline{s}$. Dunque, attraverso questa nuova notazione il numero dei gruppi di riferimento per la generica partizione P_i è data dal prodotto ${}^c\underline{S} \times {}^{\bar{c}}\underline{S} \times \underline{S}^i \times Q^i$. Come è illustrato nella tabella A.4.12 la scissione di \underline{s} nei due sottoinsiemi, consente l'attuazione dello schema C.

Tabella A.4.12 – Descrizione degli schemi di costruzione del data-set di input: definizione del numero di modalità della variabile POP_PIAN e del numero di variabili X_j

SCHEMA		Numero delle modalità della variabile POP_PIAN		Numero di variabili X_j per ogni variabile X_t definita in P_1		Numero di variabili X_j per ogni variabile X_t definita in P_i		Numero di variabili X_j per ogni variabile X_t definita in P_i
A		1	\Rightarrow	$\underline{S} \times \underline{S}^1 \times Q^1$	\cdot	$\underline{S} \times \underline{S}^i \times Q^i$	\cdot	$\underline{S}^i \times Q^i$
B		\underline{S}	\Rightarrow	$\underline{S}^1 \times Q^1$	\cdot	$\underline{S}^i \times Q^i$	\cdot	$\underline{S}^i \times Q^i$
C	Due alternative	${}^c\underline{S}$	\Rightarrow	${}^{\bar{c}}\underline{S} \times \underline{S}^i \times Q^i$	\cdot	${}^{\bar{c}}\underline{S} \times \underline{S}^i \times Q^i$	\cdot	${}^{\bar{c}}\underline{S} \times \underline{S}^i \times Q^i$
		${}^{\bar{c}}\underline{S}$		${}^c\underline{S} \times \underline{S}^i \times Q^i$	\cdot	${}^c\underline{S} \times \underline{S}^i \times Q^i$	\cdot	${}^c\underline{S} \times \underline{S}^i \times Q^i$

Dalla tabella si evidenziano alcune considerazioni già espresse a conclusione dell'esempio A.4.1: in primo luogo lo schema B è inapplicabile quando \underline{s} è un insieme vuoto (o, in altri termini lo schema B coincide con lo schema A); in secondo luogo lo schema C è inapplicabile quando \underline{s} contiene una sola variabile (o, in altri termini lo schema C coincide con lo schema B).

A.4.2 Costruzione dei gruppi di riferimento: caso II

Gruppi di riferimento nel caso di sottopopolazioni pianificate ottenute non marginalizzando la variabile di stratificazione multivariata e con variabili qualitative ausiliarie X di tipo non dicotomico

Gli schemi illustrati nella tabella A.4.12 non comprendono tutti i tipi di partizioni in gruppi di riferimento e tutti i tipi di variabili ausiliarie che

possono essere state utilizzate per definire lo stimatore di ponderazione vincolata che ha generato i coefficienti finali di riporto di input. Infatti, nel descrivere l'impostazione del *data-set* di input si è fatto riferimento a due ipotesi restrittive che non sempre si verificano nella pianificazione di una strategia di campionamento: la prima ipotesi prevede che il processo di aggregazione degli strati per definire le sottopopolazioni pianificate avvenga marginalizzando rispetto ad una o più variabili che individuano gli stessi strati; la seconda suppone che le variabili qualitative x siano dicotomiche, del tipo presenza/assenza, sì/no, 0/1, ecc..

Di seguito sono illustrati i passi necessari per impostare il *data-set* di input quando le ipotesi precedenti non sono proprie della strategia campionaria adottata dall'utente.

Per comprendere quali sono le implicazioni che intervengono quando non si verifica la prima ipotesi è utile considerare il seguente esempio:

Esempio A.4.2:

Sia dato un disegno campionario in cui la stratificazione avviene su una variabile multivariata ottenuta dalle variabili sesso (2 modalità; uomini - U; donne - D) e classe di età (4 modalità: 0-14 anni; 15-34 anni; 35-54 anni; 55 anni e oltre). Su tale stratificazione si può effettuare un primo tipo di aggregazione degli strati marginalizzando sulla classe di età e formando, pertanto, due gruppi di strati: il primo identificato dalla modalità U, il secondo dalla modalità D.

Con questa stratificazione la strategia campionaria potrebbe, tuttavia, presentare un secondo tipo di aggregazione degli strati che coinvolge l'unione di alcune modalità all'interno di una variabile che identifica gli strati senza procedere alla marginalizzazione rispetto ad una specifica variabile. Ciò avviene, ad esempio, aggregando gli strati identificati dalle modalità 0-14 anni e 15-34 anni della variabile classe di età, ottenendo, dunque, sei gruppi di strati, identificati esattamente da: U e 0-34 anni, D e 0-34 anni, U e 35-54 anni, D e 35-54 anni, U e 55 anni e oltre, D e 55 anni e oltre.

L'aggregazione degli strati che non prevede una marginalizzazione rispetto a variabili che identificano gli strati stessi non presenta particolari problemi dal punto di vista operativo. Bisogna, tuttavia, distinguere due casi:

- *il primo prevede che la procedura di aggregazione degli strati è la stessa su tutte le partizioni considerate;*

- *il secondo permette di avere differenti procedure di aggregazione che cambiano al cambiare delle partizioni. Riprendendo l'esempio, la strategia campionaria potrebbe presentare una prima partizione ottenuta aggregando degli strati con classe di età 0-14 anni e 15-34 anni e una seconda partizione in cui si aggregano fra loro gli strati con modalità 0-14 anni e 15-34 anni e gli strati con modalità 35-54 anni e 55 anni e oltre. In tutti i casi le aggregazioni avvengono per strati che presentano la stessa modalità della variabile sesso.*

Facendo riferimento alla suddivisione delle variabili, in variabili che definiscono gli strati e variabili di post-stratificazione, necessaria per impostare l'archivio di input, si deve procedere in due modi differenti per i due casi:

- *nel primo caso si sostituisce la variabile in cui avvengono le aggregazioni, con una nuova variabile le cui modalità sono aggregazioni delle modalità di quella originale. Così, se l'aggregazione degli strati è quella presentata nell'esempio e se questo criterio di aggregazione si ripete in tutte le partizioni previste dalla strategia campionaria, la variabile classe di età con quattro modalità (0-14 anni;15-34 anni;35-54 anni;55 anni e oltre) è sostituita nella definizione del data-set di input con una nuova variabile che presenta tre modalità (0-34 anni;35-54 anni;55 anni e oltre);*
- *nel secondo caso l'originale variabile di stratificazione non viene considerata nella formazione del data-set, mentre sono prese in considerazione tante nuove variabili di post-stratificazione per quante sono le differenti aggregazioni in strati. Ad esempio, considerando il punto ii nella costruzione dell'archivio di input si deve escludere la variabile classe di età come variabile che definisce gli strati e si devono inserire una prima nuova variabile di post-stratificazione con tre modalità (0-34;35-54;55 e oltre) e una seconda nuova variabile di post-stratificazione con due modalità (0-34;35 e oltre).*

Considerando ora il caso in cui le variabili qualitative inserite nel processo di calibrazione non sono dicotomiche (presenza/assenza; 0/1 ecc.) è necessario operare una loro preventiva trasformazione nella forma detta disgiuntiva completa.

Sia data per esempio la variabile “titolo di studio” con quattro modalità: “licenza elementare”; “licenza media”; “diploma di scuola superiore”; “laurea universitaria”. In questo caso la forma disgiuntiva completa della variabile definisce le seguenti quattro variabili dicotomiche: “il titolo di

studio è la licenza elementare con modalità si/no”; “il titolo di studio è la licenza media con modalità si/no”; “il titolo di studio è diploma di scuola superiore con modalità si/no”; “il titolo di studio è la laurea universitaria con modalità si/no”.

Sulla base di queste quattro variabili saranno definite successivamente le variabili X_j secondo l’opportuno schema di impostazione del *data-set* di input.

A.5 Presentazione sintetica degli errori di campionamento mediante modelli regressivi

A.5.1. Introduzione

Una informazione completa sul livello di precisione dei risultati prodotti da un'indagine campionaria richiederebbe la specificazione degli errori campionari di tutte le stime pubblicate. Tuttavia, le indagini su larga scala prodotte dai principali centri di diffusione statistica a livello nazionale ed internazionale sono caratterizzate da strategie campionarie complesse - basate su disegni campionari ad uno o più stadi di selezione, con stratificazione delle unità primarie, selezione delle unità con probabilità variabili e senza reimmissione, e utilizzano stimatori che sono funzioni non lineari dei dati campionari - e da un numero estremamente elevato di stime prodotte. Risulterebbe, quindi, oneroso e di difficile attuazione, per limiti di tempo e di costi di elaborazione, pubblicare per ciascuna stima il corrispondente errore campionario. Inoltre, le tavole di pubblicazione sarebbero appesantite e di non facile consultazione per l'utente finale.

Tali difficoltà hanno portato allo studio di alcuni metodi approssimati che agevolano notevolmente il calcolo degli errori campionari ed idonei modelli che consentono di esporre in forma concisa i suddetti errori. Tali modelli si possono suddividere in due tipi, a seconda della metodologia utilizzata: quella dei *modelli regressivi* e quella basata sull'*effetto del disegno di campionamento* (o *deft, design effect*) (Verma, Scott e Muirheartaigh, 1980; Verma, 1982; Wolter, 1985). La metodologia che è implementata dal software è quella dei modelli regressivi ed è fondata sulla determinazione di

una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento.

L'approccio utilizzato nel software per la costruzione dei modelli regressivi è differente a seconda che le stime di interesse siano:

- (i) stime di frequenze assolute o relative, riferite alle modalità di una variabile qualitativa, oppure alle classi formate in base ad una variabile quantitativa; esempi di stime di questo tipo sono:
 - la stima del numero totale di individui della popolazione che risultano occupati, oppure la stima del numero totale di individui appartenenti alla classe di età [10-12) anni;
 - la stima del numero totale di imprese della popolazione che producono un dato tipo di prodotto, oppure la stima del numero totale di imprese che appartengono alla classe dimensionale [1-3) addetti;
- (ii) stime di totali di variabili quantitative; esempi di stime di questo tipo sono:
 - il valore monetario complessivo delle spese effettuate dalle famiglie italiane nel mese di dicembre, oppure il numero totale di viaggi di lavoro effettuati dagli individui della popolazione italiana nel primo trimestre dell'anno;
 - il totale degli addetti che lavorano nelle imprese italiane, oppure il totale degli investimenti effettuati da tali imprese.

Per le stime del tipo (i) è possibile utilizzare modelli regressivi che hanno un fondamento teorico, secondo cui gli errori relativi delle stime di frequenze sono espressi da una funzione decrescente al crescere dei valori delle stime stesse. Per le stime del tipo (ii), invece, il problema è piuttosto complesso, dal momento che non è stata ancora elaborata un'adeguata base teorica per l'interpolazione degli errori campionari delle stime in questione. L'approccio adottato per trattare il caso di variabili quantitative è pertanto di tipo empirico ed è fondato sull'evidenza sperimentale che l'errore assoluto di un totale è una funzione crescente del totale stesso.

Nel seguito del paragrafo verranno descritti separatamente i modelli regressivi adottati per le stime del tipo (i) ed (ii). Una trattazione appro-

fondita degli argomenti di seguito trattati è contenuta anche nel lavoro di Russo (1987).

A.5.2. Caratteristiche generali del metodo

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichino rispettivamente con $V(\hat{Y}_\omega)$ e con $\sigma(\hat{Y}_\omega) = \sqrt{V(\hat{Y}_\omega)}$, la varianza e l'errore di campionamento della stima \hat{Y}_ω del generico parametro di interesse Y_ω ($\omega=1, \dots, \Omega$); si indichino, inoltre, con

$$\varepsilon^2(\hat{Y}_\omega) = \frac{V(\hat{Y}_\omega)}{Y_\omega^2}, \quad \varepsilon(\hat{Y}_\omega) = \frac{\sigma(\hat{Y}_\omega)}{Y_\omega}$$

le corrispondenti quantità relative.

Denotato con $G = \{\hat{Y}_\omega, (\omega=1, \dots, \Omega)\}$ l'insieme delle stime di interesse, l'ipotesi fondamentale alla base del metodo dei modelli regressivi è quella che, nell'ambito dell'insieme G , l'errore campionario relativo, $\varepsilon(\hat{Y}_\omega)$, oppure la varianza campionaria relativa, $\varepsilon^2(\hat{Y}_\omega)$, dipendono soltanto dall'ampiezza del parametro Y_ω . Ad esempio è possibile definire un legame funzionale che lega la varianza relativa $\varepsilon^2(\hat{Y})$ di una stima \hat{Y} , con il corrispondente valore del parametro di interesse Y mediante la seguente relazione funzionale:

$$\varepsilon^2(\hat{Y}) = f(Y, \alpha_1, \dots, \alpha_q, u) \quad (\text{A.5.1})$$

in cui $\alpha_1, \dots, \alpha_q$ sono dei parametri incogniti e u è un errore casuale.

In pratica la precedente relazione viene sostituita dall'analoga relazione operativa

$$\hat{\varepsilon}^2(\hat{Y}) = f(\hat{Y}, \alpha_1, \dots, \alpha_q, u) \quad (\text{A.5.2})$$

in cui

$$\hat{\varepsilon}^2(\hat{Y}) = \frac{\hat{V}(\hat{Y})}{\hat{Y}^2}$$

La stima dei parametri $\alpha_1, \dots, \alpha_q$ si ottiene adattando il modello (A.5.2) ad una nuvola di punti $(\hat{Y}_\omega, \hat{\varepsilon}^2(\hat{Y}_\omega))$ formata da un sotto insieme,

$G' = \{\hat{Y}_\omega, (\omega = 1, \dots, \Omega')\}$ di numerosità Ω' ($\Omega' \leq \Omega$), delle stime appartenenti all'insieme G e dalle corrispondenti varianze relative $\{\varepsilon^2(\hat{Y}_\omega), (\omega = 1, \dots, \Omega')\}$.

Si perviene, pertanto, al seguente modello stimato

$$\hat{\varepsilon}^2(\hat{Y}) = f(\hat{Y}, \hat{\alpha}_1, \dots, \hat{\alpha}_q, e) \quad (\text{A.5.3})$$

in cui $\hat{\alpha}_1, \dots, \hat{\alpha}_q$ indicano rispettivamente le stime dei parametri incogniti $\alpha_1, \dots, \alpha_q$ ed e rappresenta il residuo ottenuto come

$$e = \hat{\varepsilon}^2(\hat{Y}) - \hat{\varepsilon}^2(\hat{Y})$$

essendo $\hat{\varepsilon}^2(\hat{Y})$ il corrispondente valore stimato della varianza relativa della stima \hat{Y} , ottenuto attraverso la relazione

$$\hat{\varepsilon}^2(\hat{Y}) = f(\hat{Y}, \hat{\alpha}_1, \dots, \hat{\alpha}_q) .$$

Per ciascuna stima appartenente all'insieme G è possibile, quindi, determinare una stima della corrispondente varianza relativa mediante la relazione

$$\hat{\varepsilon}^2(\hat{Y}_\omega) = f(\hat{Y}_\omega, \hat{\alpha}_1, \dots, \hat{\alpha}_q) \quad (\text{A.5.4})$$

A partire dalla (A.5.4) è possibile, poi, ottenere l'errore relativo ed assoluto, espressi rispettivamente da

$$\hat{\varepsilon}(\hat{Y}_\omega) = \sqrt{f(\hat{Y}_\omega, \hat{\alpha}_1, \dots, \hat{\alpha}_q)} \quad (\text{A.5.5})$$

$$\hat{\sigma}(\hat{Y}_\omega) = \hat{\varepsilon}(\hat{Y}_\omega) \hat{Y}_\omega \quad (\text{A.5.6})$$

Al fine di permettere il calcolo degli errori campionari delle stime pubblicate, mediante il metodo appena descritto, nei volumi in cui vengono presentati i risultati di un indagine campionaria viene riportata, usualmente, una tabella del seguente tipo:

Tabella A14: coefficienti stimati del modello (A.5.2) e grado di adattamento del modello a livello totale e per ciascun dominio di studio

	Coefficienti stimati del modello			Indice di determinazione %
Totale	$\hat{\alpha}_1$	$\hat{\alpha}_q$	R^2
Dominio di studio 1	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{q,1}$	R_1^2
...
Dominio di studio D	$\hat{\alpha}_{1,D}$	$\hat{\alpha}_{q,D}$	R_D^2

in cui, con riferimento a ciascun dominio di studio d ($d=1,\dots,D$) e per il totale della popolazione sono contenuti i valori dei coefficienti stimati $\hat{\alpha}_1,\dots,\hat{\alpha}_q$. Al fine di documentare il grado di rappresentatività degli errori campionari stimati in base al modello (A.5.2), in tale tabella viene riportato, con riferimento a ciascun dominio di studio d , il coefficiente di determinazione R_d^2 che rappresenta il grado di adattamento della funzione interpolata alla nuvola di punti

$$\left(\hat{y}_{\omega,d}, \hat{\epsilon}^2(\hat{y}_{\omega,d}) \right).$$

Poiché per gli utenti non statistici il calcolo degli errori campionari mediante i modelli interpolati (A.5.5) può risultare di non facile utilizzo, si affianca generalmente alla tabella A14 una tabella che permette una valutazione più agevole degli errori campionari delle stime pubblicate, anche se conduce a risultati meno precisi. La suddetta tabella, che viene presentata con riferimento a ciascun dominio di studio, è del seguente tipo:

Tabella A15: valori interpolati degli errori relativi in corrispondenza ad alcuni valori tipici prefissati delle stime, a livello totale e per ciascun dominio di studio

Dominio di studio 1			Dominio di studio D		Totale	
Livelli di stima prefissati	Errori relativi interpolati	Livelli di stima prefissati	Errori relativi interpolati	Livelli di stima prefissati	Errori relativi interpolati
$\hat{Y}_{1,1}^*$	$\hat{\varepsilon}(\hat{Y}_{1,1}^*)$	$\hat{Y}_{1,D}^*$	$\hat{\varepsilon}(\hat{Y}_{1,D}^*)$	\hat{Y}_1^*	$\hat{\varepsilon}(\hat{Y}_1^*)$
.
.
$\hat{Y}_{k,1}^*$	$\hat{\varepsilon}(\hat{Y}_{k,1}^*)$	$\hat{Y}_{k,D}^*$	$\hat{\varepsilon}(\hat{Y}_{k,D}^*)$	\hat{Y}_k^*	$\hat{\varepsilon}(\hat{Y}_k^*)$
.
.
$\hat{Y}_{K,1}^*$	$\hat{\varepsilon}(\hat{Y}_{K,1}^*)$	$\hat{Y}_{K,D}^*$	$\hat{\varepsilon}(\hat{Y}_{K,D}^*)$	\hat{Y}_K^*	$\hat{\varepsilon}(\hat{Y}_K^*)$

Nella prima e nella seconda colonna della tabella A15 sono riportati rispettivamente:

- alcuni particolari livelli di stima; così, ad esempio, nel caso dell'indagine Multiscopo, per la stima di frequenze assolute riferite alle famiglie si utilizzano i seguenti livelli di stima: 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 750, 1000, 2000, 3000, 4000, 5000, 7500, 15000, 20000 e 25000 migliaia, sia con riferimento a ciascun dominio di stima considerato che con riferimento al totale popolazione. Poiché in tal caso la colonna relativa alle stime $\hat{Y}_{k,d}^*$ ($k=1, \dots, K$), dove K è l'indice del parametro d'interesse, è sempre la stessa per tutti i domini di studio d ed anche per il totale popolazione, la struttura della tabella A15 sopra riportata viene leggermente modificata in quanto la colonna relativa alle stime $\hat{Y}_{k,d}^*$ viene riportata nella tabella una sola volta per tutti i domini anziché per ciascun dominio separatamente;
- i corrispondenti valori dell'errore relativo riferiti ad un particolare dominio di studio d ed al totale popolazione, ottenuti attraverso il

modello (A.5.5) ponendo rispettivamente $\hat{Y}_{k,d}^* = \hat{Y}_\omega$ (per $d=1, \dots, D$)
e $\hat{Y}_k^* = \hat{Y}_\omega$.

Il software costruisce, su richiesta dell'utente, entrambe le tabelle A14 e A15 sopra descritte; in particolare per quanto riguarda la definizione dei valori $\hat{Y}_{k,d}^*$ ($k=1, \dots, K$ e $d=1, \dots, D$) della tabella A15, si opera nel seguente modo:

- per ciascun dominio d si calcola il totale popolazione T_d , ottenuto come somma dei pesi finali (COEFFIN) delle unità elementari appartenenti al dominio stesso;
- si calcola il totale popolazione, T , ottenuto come somma dei pesi finali (COEFFIN) di tutte le unità elementari intervistate;
- si definiscono alcuni valori tipici prefissati di stime di frequenze percentuali P_k^* (per $P_k^* = 0,1; 0,5; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 15; 20; 25; 30; 35; 40; 45; 50$)
- si calcolano i corrispondenti valori delle stime di frequenze assolute mediante le seguenti formule

$$\hat{Y}_{k,d}^* = P_k^* T_d \quad (k=1, \dots, K \text{ e } d=1, \dots, D)$$

e

$$\hat{Y}_k^* = P_k^* T \quad ,$$

riferite rispettivamente al generico dominio d ed al totale popolazione.

Il calcolo dell'errore relativo corrispondente alla generica stima $\hat{Y}_{\omega,d}$ appartenente all'insieme G_d delle stime pubblicate con riferimento al dominio d può essere ricavato, a partire dalla tabella A15, in base ad uno dei seguenti metodi:

- (1) il primo metodo consiste nell'individuare, sulla colonna della tabella A15 riferita al dominio d , il livello di stima che più si avvicina alla stima di interesse $\hat{Y}_{\omega,d}$ e nel considerare come errore relativo il valore che si trova sulla stessa riga della seconda colonna della tabella riferita a detto dominio di studio;
- (2) nel secondo metodo, l'errore campionario della stima $\hat{Y}_{\omega,d}$ si ricava mediante la seguente espressione

$$\hat{\hat{e}}(\hat{Y}_{\omega,d}) = \hat{\hat{e}}(\hat{Y}_{k-1,d}^*) - \frac{\hat{\hat{e}}(\hat{Y}_{k-1,d}^*) - \hat{\hat{e}}(\hat{Y}_{k,d}^*)}{\hat{Y}_{k-1,d}^* - \hat{Y}_{k,d}^*} (\hat{Y}_{\omega,d} - \hat{Y}_{k-1,d}^*)$$

dove $\hat{Y}_{k-1,d}^*$ e $\hat{Y}_{k,d}^*$ sono i valori delle stime, riportati nella prima colonna della tabella A15 riferita al dominio d , entro i quali è compresa la stima di interesse $\hat{Y}_{\omega,d}$ ed $\hat{\hat{e}}(\hat{Y}_{k-1,d}^*)$ e $\hat{\hat{e}}(\hat{Y}_{k,d}^*)$ sono i corrispondenti errori relativi letti sulla seconda colonna della tabella, sempre riferita al dominio d .

E' importante sottolineare il fatto che il metodo dei modelli regressivi richiede il calcolo degli errori relativi su un sottoinsieme di stime di dimensione molto minore rispetto a quella dell'insieme G_d e tale metodo, pertanto, costituisce una semplificazione e una riduzione dei costi notevole rispetto al criterio di specificare accanto ad ogni stima pubblicata il corrispondente errore di campionamento. Nel caso delle stime di frequenze assolute per l'adattamento del modello, si presceglie generalmente per ciascun dominio di stima e per il totale popolazione un sottoinsieme di circa 40 stime di interesse distribuito in modo da coprire uniformemente l'intero campo di variabilità delle stime oggetto di pubblicazione.

A.5.3. Il caso delle stime di frequenze

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichi con

$$Y = \sum_{i=1}^N Y_i \quad (\text{A.5.7})$$

il numero totale di unità della popolazione che possiedono una data caratteristica di interesse, in cui: Y_i è una variabile indicatrice pari ad uno se l'unità i -esima della popolazione presenta il carattere di interesse e zero altrimenti; N indica la numerosità totale della popolazione di interesse. Sia inoltre

$$\hat{Y} = \sum_{i=1}^n K_i Y_i \quad (\text{A.5.8})$$

una stima corretta del parametro Y in cui

$$W_i = \frac{1}{\pi_i}$$

è il peso diretto assegnato alla i -esima unità campionaria ottenuto in base

al disegno campionario complesso adottato e π_i rappresenta la probabilità di inclusione nel campione dell'unità *i-esima*.

La varianza campionaria della stima \hat{Y} può essere espressa dal prodotto della varianza di un campione casuale semplice di numerosità n per la statistica *deff* (effetto del disegno di campionamento) espresso dal quadrato del *deft* (paragrafo 6.4.2). Si ha pertanto che:

$$V(\hat{Y}) = N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n} deff \quad (A.5.9)$$

essendo

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N \left(Y_i - \frac{1}{N} \sum_{i=1}^N Y_i \right)^2. \quad (A.5.10)$$

Nel caso in esame, la precedente quantità può essere riscritta come

$$\sigma^2 = P(1-P) \quad (A.5.11)$$

in cui

$$P = \frac{Y}{N}$$

In base alle espressioni (A.5.11) e (A.5.9), la varianza relativa della stima \hat{Y} può, quindi, essere espressa da

$$\varepsilon^2(\hat{Y}) = \frac{V(\hat{Y})}{Y^2} = \frac{N^2}{Y^2} \frac{N-n}{N-1} \frac{P(1-P)}{n} deff \quad (A.5.12)$$

che attraverso semplici passaggi assume la forma

$$\varepsilon^2(\hat{Y}) = \frac{1}{n} \frac{N-n}{N-1} \left[\frac{N}{Y} - 1 \right] deff. \quad (A.5.13)$$

Ponendo

$$A = \frac{N}{n} \frac{(N-n)}{N-1} \quad (A.5.14)$$

si ottiene infine

$$\varepsilon^2(\hat{Y}) = -\frac{A}{N} deff + A deff \frac{1}{Y} \quad (A.5.15)$$

Sotto l'ipotesi che il *deff* sia costante (o approssimativamente tale), nell'ambito di un determinato insieme G di stime di frequenze assolute, è possibile formulare un modello regressivo del tipo (A.5.2) per stimare l'errore campionario delle stime appartenenti a tale insieme. In base alla (A.5.15) si ha, quindi, che

$$\varepsilon^2(\hat{Y}) = \alpha_1 + \frac{\alpha_2}{Y} + u \quad (\text{A.5.16})$$

E' possibile ottenere un modello alternativo al precedente modificando opportunamente la (A.5.15). Si ottiene, infatti, mediante semplici passaggi che tale relazione può essere riscritta come

$$\varepsilon^2(\hat{Y}) = -\frac{A \text{ deff}}{Y} \left(1 - \frac{Y}{N}\right). \quad (\text{A.5.17})$$

Calcolando il logaritmo di entrambi i membri della precedente relazione si ottiene

$$\log(\varepsilon^2(\hat{Y})) = \log(A \text{ deff}) - \log(Y) + \log\left(1 - \frac{Y}{N}\right) \quad (\text{A.5.18})$$

La precedente relazione non è lineare in $\log(\varepsilon^2(\hat{Y}))$ e $\log(Y)$ per la presenza del terzo termine a secondo membro, tuttavia per valori bassi del rapporto (Y/N) tale termine è trascurabile. Pertanto, sotto l'ipotesi che il *deff* sia costante nell'ambito dell'insieme di stime G , si ottiene il seguente modello alternativo

$$\log(\varepsilon^2(\hat{Y})) = \alpha_1 + \alpha_2 \log(Y) + u \quad (\text{A.5.19})$$

per stimare la varianza delle stime appartenenti all'insieme G .

Il corrispondente modello non lineare è espresso quindi da

$$\varepsilon^2(\hat{Y}) = \tilde{\alpha}_1 Y^{\tilde{\alpha}_2} \tilde{u} \quad (\text{A.5.20})$$

in cui si è posto

$$\tilde{\alpha}_1 = \text{anti log}(\alpha_1), \quad \tilde{\alpha}_2 = \alpha_2 \quad (\text{A.5.21})$$

e

$$\tilde{u} = \text{anti log}(u)$$

Una stima dei parametri α_1 e α_2 del modello (A.5.19) si ottiene, mediante il metodo dei minimi quadrati (semplici o ponderati, nel caso in cui viene rilasciata l'ipotesi di omoschedasticità), adattando il modello in oggetto ad una nuvola di punti $(\hat{Y}, \hat{\varepsilon}^2(\hat{Y}))$ formata da un sotto insieme di stime, appartenenti all'insieme G' , e dalle corrispondenti varianze relative.

Si perviene, in tal modo, al seguente modello stimato

$$\log(\hat{\varepsilon}^2(\hat{Y})) = \hat{\alpha}_1 + \hat{\alpha}_2 \log(\hat{Y}) + e \quad (\text{A.5.22})$$

in cui $\hat{\alpha}_1$ e $\hat{\alpha}_2$ indicano rispettivamente gli stimatori dei minimi quadrati dei parametri incogniti α_1 e α_2 ed e rappresenta il residuo ottenuto come

$$e = \hat{\varepsilon}^2(\hat{Y}) - \hat{\varepsilon}^2(\hat{Y})$$

essendo $\hat{\varepsilon}^2(\hat{Y})$ il corrispondente valore stimato della varianza relativa della stima \hat{Y} , ottenuto attraverso la relazione

$$\log(\hat{\varepsilon}^2(\hat{Y})) = \hat{\alpha}_1 + \hat{\alpha}_2 \log(\hat{Y}) \quad (\text{A.5.23})$$

E' possibile ottenere una stima dei parametri $\tilde{\alpha}_1$ e $\tilde{\alpha}_2$ del modello non lineare (A.5.20) sfruttando le relazioni (A.5.21). Si ha pertanto

$$\hat{\tilde{\alpha}}_1 = \text{anti log}(\hat{\alpha}_1), \quad \hat{\tilde{\alpha}}_2 = \hat{\alpha}_2. \quad (\text{A.5.24})$$

Si ritiene importante mettere in luce il fatto che gli stimatori dei minimi quadrati $\hat{\alpha}_1$ e $\hat{\alpha}_2$ sono stimatori non distorti dei rispettivi parametri α_1 e α_2 mentre gli stimatori $\hat{\tilde{\alpha}}_1$ e $\hat{\tilde{\alpha}}_2$, del corrispondente modello non lineare, non godono della proprietà di correttezza con riferimento ai parametri $\tilde{\alpha}_1$ e $\tilde{\alpha}_2$. L'applicazione del metodo dei minimi quadrati a funzioni linearizzate dei parametri viene spesso effettuata per comodità di calcolo, poiché i metodi di stima non lineare sono più complessi.

E' possibile utilizzare il modello (A.5.23) anche per la presentazione sintetica di stime di frequenze relative. Infatti, per ogni stima di frequenza assoluta, \hat{Y} , a cui corrisponde una stima della frequenza relativa \hat{P} , vale la ben nota relazione

$$\hat{\varepsilon}^2(\hat{P}) = \hat{\varepsilon}^2(\hat{Y}).$$

In base al modello (A.5.23) è possibile, quindi, scrivere

$$\log(\hat{\epsilon}^2(\hat{P})) = \log(\hat{\epsilon}^2(\hat{Y})) = \hat{\alpha}_1 + \hat{\alpha}_2 \log(\hat{Y}) \quad (\text{A.5.25})$$

A conclusione di questo paragrafo è importante ricordare che il software Genesees utilizza due modelli per la presentazione sintetica degli errori campionari. Il primo modello (modello 1) presentato dal software nella STAMPA 5 (*paragrafo 6.4.5*), può essere utilizzato nel caso si voglia effettuare la presentazione sintetica degli errori campionari per le stime di frequenze, il modello adattato dal software per questo caso è il modello (A.5.23) sopra riportato. In particolare nella stampa 5 vengono presentate due tabelle. La prima tabella (stampa 5a) ha una struttura analoga alla tabella A14 descritta nel *paragrafo 4.5.2*; in essa vengono riportati, per ciascun dominio di studio (detto dominio pianificato) ed al livello della popolazione totale, i valori stimati dei parametri α_1 e α_2 e l'indice di determinazione R^2 % del modello (A.5.23). In tale tabella i parametri α_1 e α_2 vengono rispettivamente indicati con i simboli "A" e "B". La seconda tabella (stampa 5b) ha una struttura analoga alla tabella A15 descritta nel *paragrafo 4.5.2*. E' importante sottolineare il fatto che, per ciascun dominio di stima, il software effettua l'adattamento del modello (A.5.23) ad una nuvola di punti definita in base a tutte le stime per le quali l'utente ha richiesto il calcolo degli errori campionari nella fase di lancio della procedura. Tale nuvola di punti è quindi definita dall'utente nella fase di lettura del *data-set* di input ed, in particolare, dipende dalla scelta delle variabili di interesse e delle sottoclassi. Il grado di adattamento del modello (A.5.23) risulta generalmente alto, tuttavia un basso valore dell'indice di determinazione R^2 % può essere dovuto alla presenza di alcuni valori *outlier* nella nuvola di punti considerata. Tale circostanza è essenzialmente legata alla presenza di alcune stime per le quali non è valida l'ipotesi di *deff* costante. In tale circostanza al fine di migliorare l'adattamento del modello ai dati occorre effettuare le seguenti operazioni, con riferimento a ciascun dominio di studio in cui si osserva un basso indice di determinazione R^2 %:

- si individuano le stime a cui sono associati valori della statistica *deft* (pari alla radice quadrata del *deff*) molto al di sotto oppure molto al di sopra del *deft* medio calcolato su tutte le stime del dominio; que-

sta operazione può essere svolta leggendo sulla STAMPA 2 (paragrafo 6.4.2) i *deft* delle differenti stime e confrontando tali *deft* con il corrispondente *deft* medio presentato nella STAMPA 6 (paragrafo 6.4.6). L'operazione di individuazione dei valori di *outlier* può essere anche facilitata confrontando il grafico dei valori osservati e dei valori interpolati (in base al modello A.5.23) degli errori relativi, corrispondenti alle differenti stime di interesse;

- si rilancia nuovamente la procedura eliminando le stime che presentano dei valori di *outlier* della statistica *deft*; oppure si eliminano direttamente tali stime dal data set TOTALE e si richiede nuovamente la stampa dei modelli.

Come si è detto, per l'adattamento del modello di scelgono, generalmente, circa 40, 50 stime di interesse che si distribuiscono uniformemente sull'intero campo di variazione delle stime pubblicate.

Un basso grado di adattamento del modello può essere anche determinato da un numero eccessivamente elevato di stime considerate; in tal caso, infatti, aumenta la possibilità che tra le stime considerate si trovino alcuni valori di outlier.

A.5.4. Il caso delle stime di totali di variabili quantitative

Si supponga di aver effettuato un'indagine basata su un disegno campionario complesso e si indichi con Y , espresso mediante la formula (A.5.7), il totale della variabile quantitativa Y in cui Y_i rappresenta il valore assunto da detta variabile con riferimento alla i -esima unità della popolazione di interesse; sia, inoltre \hat{Y} , espresso mediante la formula (A.5.8), una stima corretta del parametro Y .

Nel caso in esame, a partire dalla (A.5.9), sfruttando le seguenti espressioni

$$\sigma^2 = \frac{1}{N} \left(\sum_{i=1}^N Y_i^2 - \frac{Y^2}{N} \right), \quad (\text{A.5.26})$$

$$\left(\sum_{i=1}^N Y_i \right)^2 = \sum_{i=1}^N Y_i^2 + 2 \sum_{i=1}^N \sum_{i' > i}^N Y_i Y_{i'} \quad (\text{A.5.27})$$

la varianza della stima \hat{Y} può essere espressa come

$$V(\hat{Y}) = \left[\frac{(N-1)}{N} AY^2 - 2A \left(\sum_{i=1}^N \sum_{i' > i}^N Y_i Y_{i'} \right) \right] deff \quad (A.5.28)$$

in cui A è dato dalla (A.5.14); passando alla varianza relativa si ottiene quindi

$$\varepsilon^2(\hat{Y}) = \left[\frac{(N-1)}{N} A - \frac{2A}{Y^2} \left(\sum_{i=1}^N \sum_{i' > i}^N Y_i Y_{i'} \right) \right] deff \quad (A.5.29)$$

Sotto l'ipotesi che il $deff$ sia costante (o approssimativamente) nell'ambito di un determinato insieme G di stime, è possibile formulare un modello regressivo del tipo (A.5.2) per stimare l'errore campionario delle stime appartenenti a tale insieme, che, in base alla (A.5.29), può essere espresso da

$$\varepsilon^2(\hat{Y}) = \alpha_1 + \frac{\alpha_2}{Y^2} + u \quad (A.5.30)$$

Per tenere conto della presenza del termine $\sum_{i=1}^N \sum_{i' > i}^N Y_i Y_{i'}$ a secondo membro della (A.5.29) è possibile introdurre nel modello l'ipotesi di eteroschedasticità; pertanto con riferimento alle stime appartenenti all'insieme G , tale ipotesi è espressa da

$$E(u_\omega^2) = \sigma_\omega^2 = \sigma^2 f \left(\sum_{i=1}^N \sum_{i' > i}^N Y_{\omega,i} Y_{\omega,i'} \right) \quad (\omega = 1, \dots, \Omega) \quad (A.5.31)$$

In presenza dell'ipotesi di eteroschedasticità, una stima efficiente e corretta dei parametri α_1 e α_2 del modello (A.5.30) è ottenuta in base al metodo dei minimi quadrati ponderati; per l'applicazione di tale metodo, tuttavia, sarebbe necessario conoscere le varianze σ_ω^2 ($\omega = 1, \dots, \Omega$), oppure disporre di una loro stima. La stima delle varianze σ_ω^2 può comportare, tuttavia, un aumento notevole delle difficoltà di calcolo.

Per le ragioni sopra esposte si ricorre spesso a modelli empirici che mostrano un buon adattamento ai dati osservati. Un modello empirico, che usualmente conduce a buoni risultati, è il seguente

$$\sigma(\hat{Y}) = \alpha_1 + \alpha_2 \hat{Y} + \alpha_3 \hat{Y}^2 + u \quad (A.5.32)$$

Poiché il modello (A.5.32) è di tipo empirico, la stima dei parametri

α_1 e α_2 α_3 deve essere ottenuta in base ad una nuvola di punti formata utilizzando tutte le stime incluse nell'insieme G . Ciò è differente dalla procedura adottata nel caso delle stime di frequenze, in cui i parametri del modello vengono stimati in base ad una nuvola di punti formata da un sottoinsieme G' delle stime d'interesse. Nella situazione esaminata, infatti, la procedura adottata per le stime di frequenze non garantisce il buon adattamento del modello stesso anche alle stime dell'insieme G che non appartengono a G' .

A partire dalla (A.5.32) è possibile, quindi, stimare l'errore relativo di campionamento di una generica stima appartenente all'insieme G mediante le seguente espressione

$$\hat{\varepsilon}(\hat{Y}) = \hat{\alpha}_2 + \frac{\hat{\alpha}_1}{\hat{Y}} + \hat{\alpha}_3 \hat{Y} \quad (\text{A.5.33})$$

in cui, $\hat{\alpha}_1$, $\hat{\alpha}_2$ e $\hat{\alpha}_3$ rappresentano le stime dei corrispondenti parametri α_1 e α_2 α_3 , ottenute in base al metodo dei minimi quadrati.

Esplicitando la precedente espressione rispetto al valore della stima \hat{Y} si perviene alla seguente equazione di secondo grado:

$$\hat{\alpha}_1 + [\hat{\alpha}_2 - \hat{\varepsilon}(\hat{Y})] \hat{Y} + \hat{\alpha}_3 \hat{Y}^2 = 0 \quad (\text{A.5.34})$$

le cui radici sono espresse rispettivamente da

$$\hat{Y}_1 = \frac{-[\hat{\alpha}_2 - \hat{\varepsilon}(\hat{Y})] - \sqrt{[\hat{\alpha}_2 - \hat{\varepsilon}(\hat{Y})]^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2\hat{\alpha}_3} \quad (\text{A.5.35})$$

$$\hat{Y}_2 = \frac{-[\hat{\alpha}_2 - \hat{\varepsilon}(\hat{Y})] + \sqrt{[\hat{\alpha}_2 - \hat{\varepsilon}(\hat{Y})]^2 - 4\hat{\alpha}_1\hat{\alpha}_3}}{2\hat{\alpha}_3} .$$

Utilizzando le precedenti formule è possibile costruire una tabella alternativa (alla tabella A15 presentata nel *paragrafo* A.5.2) di presentazione sintetica degli errori di campionamento la cui struttura è mostrata nel seguente esempio.

Tabella A16: valori dei totali corrispondenti ad alcuni valori tipici prefissati degli errori relativi a livello di totale popolazione e per ciascun dominio di studio

	Valori prefissati degli errori relativi percentuali		
	ε_1^*		ε_K^*
Totale	\hat{Y}_1^*	\hat{Y}_K^*
Dominio di studio 1	$\hat{Y}_{1,1}^*$	$\hat{Y}_{K,1}^*$
...			
Dominio di studio D	$\hat{Y}_{1,D}^*$	$\hat{Y}_{K,D}^*$

In essa vengono riportati i valori delle stime \hat{Y}^* ottenuti in base alla (A.5.34), in relazione ad alcuni valori tipici prefissati dell'errore relativo percentuale. Definito, pertanto, con ε_k^* ($k=1, \dots, K$) il generico valore prefissato dell'errore relativo, sostituendo tale valore nella (A.5.35), al posto di $\hat{\varepsilon}(\hat{Y})$, è possibile ricavare il corrispondente valore della stima \hat{Y}_k^* scegliendo il valore assunto dalla corrispondente radice positiva dell'equazione (A.5.34) ottenuta mediante una delle (A.5.34).

La lettura di tale tabella indica che le stime con valori superiori a \hat{Y}_k^* presentano valori dell'errore relativo inferiori a ε_k^* , mentre le stime che assumono valori inferiori a \hat{Y}_k^* presentano valori dell'errore relativo superiori a ε_k^* . I valori di ε_k^* che vengono usualmente utilizzati per la costruzione della tabella sono 5, 10, 15, 20, 25, 30 e 35%.

Bibliografia

Brewer, K.R.V., Hanif, M., 1983, Sampling with Unequal Probabilities, Springer-Verlag, New-York.

Chen, P. P. S., 1976, *The Entity-Relationship Model. Towards a Unified View of Data*, ACM Trans. Database System 1, n. 1.

Cochran, W. G., 1977, Sampling Techniques, Wiley, New York.

Deville, J. C., Särndal, C. E., 1992, *Calibration Estimators in Survey Sampling*, Journal of the American Statistical Association, vol. 87, pp. 367-382.

De Vitiis, C., Pagliuca, D., 2003, *La presentazione sintetica degli errori campionari e l'analisi grafica degli outlier nel software Genesee*, Atti del Convegno Intermedio "Analisi Statistica Multivariata per le scienze economico-sociali, le scienze naturali e la tecnologia" della Società Italiana di Statistica (su CD-ROM).

Falorsi, P.D., Falorsi, S., 1995, *Un metodo di stima generalizzato per le indagini sulle famiglie e sulle imprese*, Rapporto di ricerca CON.PRI, Dipartimento di Scienze Statistiche "Paolo Fortunati", Università degli Studi di Bologna, n. 13.

Falorsi, P.D., Falorsi, S., 1997, *The Italian Generalized Package for Weighting Persons and Families: Some Experimental Results with Different Non-Response Models*, Statistics in Transitions Journal of the Polish Statistical Association, vol. 3, n. 2.

Falorsi, P. D., Falorsi S., 1998, *The Italian generalized estimation package: some experimental results for estimation on households suveys with different non response mechanism*, Quaderni di Ricerca, ISTAT, n.4, pp.63-94.

Falorsi, S., Rinaldelli, C., 1998, *Un Software generalizzato per il calcolo delle stime e degli errori di campionamento*, Statistica Applicata, vol. 10, n. 2 , pp. 217-234.

Falorsi, S., Pagliuca, D., Scepi, G., 1999, *Generalised Software for Sampling Errors – GSSE*, Proceedings of the Seminar on Exchange of Technology and Know-How (ETK 99), held in Prague, Czech Republic on the 13-15 October 1999, pp. 169-175.

Falorsi, S., Pagliuca, D., Scepi, G., 2000, *Generalised Software for Sampling Errors – GSSE*, Research in Official Statistics - ROS, vol. 3, n. 2, pp. 89-108.

Horvitz, D.G., Thompson, D. J, 1952, *A Generalization of Sampling without Replacement from Finite Universe*, Journal of the American Statistical Association, vol. 47, pp. 663-685.

Kish, L., 1965, *Survey Sampling*, Wiley, New York.

Pagliuca, D., Righi, P., 2002, *Genesees v1.0*, Proceedings of the Conference CompStat 2002 – Short Communications and Posters, Berlin August 24-th to August 28th 2002 (disponibile su CD-ROM)

Pagliuca, D. (a cura di), 2004, *Genesees V.3.0., Funzione Riponderazione*, Manuale utente ed aspetti metodologici, Tecniche e Strumenti, ISTAT, n. 2. (disponibile anche su sito: Istat <http://www.istat.it>)

Russo A., 1987, *Sulla Presentazione degli Errori di Campionamento mediante Modelli. Il Metodo dei Modelli Regressivi*, Quaderni di Discussione, ISTAT, n. 87, 04.

Särndal, C.E., Swensson , B. and Wretman, J., 1989, *The weighted residual technique for estimating the variance of the general regression estimator of the finite population total*, Biometrika, vol. 76, n. 3, pp. 527-537

Särndal, C.E., Swensson, B. and Wretman, J., 1992, *Model Assisted Survey Sampling*, Springer-Verlag. New-York.

Singh, A. C., Mohl, C. A., 1996, *Understanding Calibration Estimators in Survey Sampling*, Survey Methodology, vol. 22, n. 2, pp. 107-115.

Verma, V., Scott, C., O'Muircheartaigh, C., 1980, *Sample Designs and Sampling Errors fo the Word Fertility Survey*, Journal of the Royal Statistical Society A, vol. 143, Part. 4, pp. 431-473.

Verma, V., 1982, *The Estimation and Presentation of Sampling Errors*, Technical Bulletins, World Fertility Survey, New York.

Wolter, K. M., 1985 *Introduction to variance estimation*. Springer-Verlag. New York.

Woodruff, R.S., 1971, *A Simple Method for Approximating the Variance of a Complicated Estimate*, Journal of the American Statistical Association, vol.66, n. 334, pp. 411-414.

Collana - TECNICHE E STRUMENTI

Volumi pubblicati

- 1 - 2004 **CONCORD V. 1.0 - Controllo e correzione dei dati**
Manuale utente e aspetti metodologici ●

- 2 - 2004 **GENESEES V. 3.0 - Funzione Riponderazione**
Manuale utente e aspetti metodologici ●

- 3 - 2005 **GENESEES V. 3.0 - Funzione Stime ed Errori**
Manuale utente e aspetti metodologici ●



dati forniti su floppy



dati forniti su cd-rom

