Appendice B

Disegno di campionamento e precisione dei risultati

Nicoletta Cibella, Claudia De Vitiis

1. Il disegno campionario

L'indagine sulla disabilità è stata condotta nel 2004 attraverso metodo CATI (computer assisted telephone interview) su un sottocampione dell'Indagine "Condizioni di salute e ricorso ai servizi sanitari" 1999-2000[1]. In particolare, sono state considerate le persone non anziane che, in occasione dell'indagine del 1999-2000, avevano riferito di avere limitazioni nelle abituali attività della vita quotidiana o di essere affetti da invalidità. Sono state escluse però le persone che durante l'indagine o dai controlli preliminari alle interviste sono risultate decedute, istituzionalizzate, trasferite all'estero e le persone che al momento dell'intervista hanno dichiarato limitazioni lievi, considerate non rilevanti per gli obiettivi dell'indagine. Pertanto la rilevazione è stata condotta su un campione di 4.011 individui di età compresa tra i 4 e i 67 anni. A queste persone è stato rivolto un questionario per indagare su particolari aspetti della disabilità e sul livello di integrazione sociale delle persone disabili.

A causa della mancata disponibilità alla reintervista riferita da una parte consistente del collettivo in occasione dell'indagine sulla salute 1999-2000, del tempo trascorso tra le due indagini e della difficoltà nel reperire i numeri di telefono delle famiglie che erano state intervistate in maniera diretta, la raccolta dati dell'indagine sulle persone con disabilità è stata affetta da un elevato tasso di non risposta (superiore al 50%). Risultano quindi intervistate 1.632 persone.

2. Il calcolo dei pesi e la correzione per mancata risposta

Il principio su cui è basato ogni metodo di stima campionaria è che le unità appartenenti al campione rappresentino anche le unità della popolazione che non sono incluse nel campione; tale obiettivo viene conseguito attribuendo alle unità rispondenti un peso campionario (detto anche coefficiente di riporto all'universo), che indica il numero di unità della popolazione che l'unità stessa rappresenta. Per l'indagine sulla disabilità, nel calcolo dei pesi si è dovuto tenere conto del fatto che le unità campionarie costituiscono un sottoinsieme del campione dell'indagine "Condizioni di salute e ricorso ai servizi sanitari" e che pertanto già possiedono un coefficiente di riporto all'universo, coerente con gli obiettivi dell'indagine 1999-2000. Se non ci fossero state mancate risposte, questo stesso peso sarebbe stato utilizzato anche per l'indagine sulla disabilità senza ulteriori modifiche. In presenza di mancate risposte, tuttavia, è stato necessario modificare tali pesi per far sì che le unità rispondenti rappresentassero anche le unità non rispondenti.

Nel contesto dei metodi di correzione per mancata risposta, una procedura comune è dividere l'intero campione in celle di aggiustamento (si veda in proposito Sarndal[2], 1992) sulla base di variabili ausiliarie correlate con la probabilità di risposta, per poi riponderare le unità rispondenti attraverso una stima del tasso di risposta. Ovviamente le variabili ausiliarie devono essere note sia per i rispondenti che per i non rispondenti; in questo modo è possibile correggere il peso campionario iniziale delle unità rispondenti in modo che esse rappresentino anche le non rispondenti appartenenti alla stessa cella, con lo scopo di ridurre la distorsione causata dalla non risposta.

In questo caso tutte le variabili raccolte nell'indagine "Condizioni di salute e ricorso ai servizi sanitari" 1999-2000 erano utilizzabili come variabili ausiliarie per la definizione di correttori per mancata

[1] L'Indagine Multiscopo "Condizioni di salute e ricorso ai servizi sanitari" è basata su un campione in due stadi: le unità primarie di campionamento sono i Comuni mentre le unità finali sono le famiglie, scelte casualmente dalle anagrafi dei comuni estratti; ogni componente della famiglia campione è stato intervistato con metodo PAPI (paper and pencil interview). Il campione del 1999-2000 era costituito da circa 1.463 Comuni e 60.000 famiglie. Per dettagli sul disegno campionario e su altri aspetti dell'Indagine Istat "Condizioni di salute e ricorso ai servizi sanitari" anni 1999-2000 si faccia riferimento al volume "Le Condizioni di salute della popolazione – Anni 1999-2000" Collana: Informazioni, n. 12 Anno di edizione: 2002.

[2] Sarndal C.E., Swensson B., Wretman J. (1992) *Model assisted survey sampling,* Springer Verlag, New York.

risposta ma, in base ad analisi preliminari dei dati, solo alcune di esse sono state scelte per la loro maggiore correlazione con il fenomeno della non risposta. Le variabili considerate, quindi, sono: la presenza del numero di telefono alla prima indagine, l'area di disabilità, la riduzione dell'autonomia, il sesso, il grado di difficoltà nella vita di ogni giorno, la gravità della disabilità, la classe di età, la dimensione della famiglia, la ripartizione geografica e il livello di istruzione.

Per costruire le celle di aggiustamento sono stati utilizzati sia il modello logistico sia metodi basati sugli alberi di classificazione, allo scopo di individuare le variabili maggiormente esplicative della propensione a rispondere. Dai risultati emersi e in base anche al numero di unità rispondenti incluse in ogni cella, si è deciso di costruire delle celle di aggiustamento definite solamente dalle modalità della variabile "*gravità della disabilità*"[3] . Quindi, alle unità rispondenti all'indagine sulla disabilità è stato assegnato un peso dato dal prodotto del peso campionario ad esse attribuito per l'indagine "Condizioni di salute e ricorso ai servizi sanitari", corretto per mancata risposta utilizzando l'inverso del tasso di risposta nelle celle di aggiustamento. Infine è stata effettuata una post-stratificazione per sesso e tre classi di età per fare in modo che i 1.632 rispondenti riproducessero la medesima struttura per sesso ed età del campione completo dell'indagine Salute.

3. Valutazione del livello di precisione delle stime

Le principali statistiche di interesse per valutare la variabilità campionaria delle stime prodotte da un'indagine sono l'errore di campionamento assoluto e l'errore di campionamento relativo (o coefficiente di variazione). Indicando con $\hat{V}ar(\hat{Y}_d)$ la stima della varianza della generica stima $\hat{Y}_d$, la stima dell'errore di campionamento assoluto di $\hat{Y}_d$ si può ottenere mediante la seguente espressione

$$\hat{\sigma}(\hat{Y}_d) = \sqrt{\hat{V}ar(\hat{Y}_d)}; \qquad (1)$$

la stima dell'errore di campionamento relativo di $\hat{Y}_d$ è invece definita dall'espressione

$$\hat{\varepsilon}(\hat{Y}_d) = \frac{\hat{\sigma}(\hat{Y}_d)}{\hat{Y}_d} \qquad (2)$$

Come è stato descritto nel paragrafo 2, le stime prodotte dall'indagine sono state ottenute mediante una procedura complessa basata su passaggi di correzione per mancata risposta e post-stratificazione. Poiché, quindi, lo stimatore adottato non è funzione lineare dei dati campionari, per la stima della varianza $\hat{V}ar(\hat{Y}_d)$ si fatto ricorso a un metodo basato sulle replicazioni del campione. In particolare, si è utilizzato il metodo dei gruppi casuali (Sarndal, 1992), che ha consentito di ottenere una valutazione della varianza campionaria che tenga conto sia della correzione per mancata risposta, sia del passaggio di post-stratificazione.

Gli errori campionari ottenuti consentono di valutare il grado di precisione delle stime; inoltre, l'errore assoluto permette di costruire un intervallo di confidenza, che, con livello di fiducia P contiene il parametro oggetto di stima. L'intervallo viene espresso come

$$\left\{\hat{Y}_d - k_p\hat{\sigma}(\hat{Y}_d) \le Y_d \le \hat{Y}_d + k_p\hat{\sigma}(\hat{Y}_d)\right\} \qquad (3)$$

Nella (3) il valore di k_P dipende dal valore fissato per la probabilità P; ad esempio, per P=0.95 si ha k=1.96.

3.1 Presentazione sintetica degli errori campionari

Poiché a ciascuna stima corrisponde un errore campionario relativo, per consentire un uso corretto delle informazioni prodotte dall'indagine sarebbe necessario pubblicare, per ogni stima, anche il

[3] La variabile presenta tre modalità che descrivono diversi livelli di gravità. È stata costruita tenendo in considerazione le diverse combinazioni di disabilità o invalidità, il bisogno di aiuto per le esigenze della vita quotidiana, la costrizione a letto, su una sedia o in casa. Il livello di gravità è valutato in base al numero di sfere di autonomia funzionale compromesse.

corrispondente errore di campionamento relativo. Tuttavia, non è possibile pubblicare tutti gli errori di campionamento delle stime fornite e non sarebbero comunque disponibili gli errori delle stime non pubblicate, che l'utente può ricavare in modo autonomo. Pertanto, si fa comunemente ricorso a una presentazione sintetica degli errori relativi basata su *modelli regressivi*, fondata sulla determinazione di una funzione matematica che mette in relazione ciascuna stima con il proprio errore di campionamento. Per le stime di frequenze assolute (o relative) riferite alle modalità di variabili qualitative, è possibile utilizzare modelli che hanno un fondamento teorico, secondo cui gli errori relativi delle stime di frequenze assolute sono funzione decrescente dei valori delle stime stesse. Il modello che viene generalmente utilizzato per le stime di frequenze assolute, con riferimento al generico dominio d, è del tipo seguente:

$$log\,\hat{\varepsilon}^2({}_d\hat{Y}) = a + b\,log({}_d\hat{Y}) \qquad (4)$$

in cui i parametri a e b vengono stimati, separatamente per ogni dominio d, utilizzando il metodo dei minimi quadrati.

Il prospetto 1 riporta i valori dei coefficienti *a* e *b* e dell'indice di determinazione R^2 delle funzioni utilizzate per l'interpolazione degli errori campionari delle stime di frequenze, per totale Italia, ripartizione geografica, sesso e le tre classi di età di pubblicazione delle stime.

Sulla base delle informazioni contenute in tali prospetti, è possibile calcolare la stima dell'errore di campionamento relativo di una determinata stima $\hat{Y}_d$ mediante la formula:

$$\hat{\varepsilon}(\hat{Y}_d) = \sqrt{exp\left(a + b\,log(\hat{Y}_d)\right)} \qquad (5)$$

che si ricava facilmente dalla (4).

Se, per esempio, la stima di frequenza assoluta $\hat{Y}_d$ si riferisce agli individui dell'Italia del Nord, l'errore relativo corrispondente si ottiene introducendo nella (5) i valori dei parametri *a* e *b* riportati nel prospetto 1 in corrispondenza della ripartizione Italia del Nord (a = 6,193417, b = -0,868940).

Il prospetto 2 rende più agevole la valutazione degli errori campionari e presenta la seguente struttura: in fiancata sono elencati i valori crescenti di stima (10.000, 20.000, …, 500.000); le colonne successive contengono gli errori di campionamento relativo, per ciascun dominio di interesse, calcolati mediante l'espressione (5), corrispondenti alle stime della prima colonna.

Le informazioni contenute in tali prospetti permettono di calcolare l'errore relativo di una generica stima (di frequenza assoluta o di un totale) mediante due procedimenti che risultano di facile applicazione, anche se conducono a risultati meno precisi di quelli ottenibili mediante l'espressione (5). Il primo metodo consiste nell'individuare, nella prima colonna del prospetto, il livello di stima che più si avvicina alla stima di interesse e nel considerare come errore relativo il valore che si trova sulla stessa riga, nella colonna corrispondente al domino di riferimento. Con il secondo metodo, l'errore campionario della stima $\hat{Y}_d$ si ricava per interpolazione mediante la seguente espressione:

$$\hat{\varepsilon}(\hat{Y}_d) = \hat{\varepsilon}(\hat{Y}_d^{k-1}) - \frac{\hat{\varepsilon}(\hat{Y}_d^{k-1}) - \hat{\varepsilon}(\hat{Y}_d^{k})}{\hat{Y}_d^{k} - \hat{Y}_d^{k-1}}(\hat{Y}_d - \hat{Y}_d^{k-1}) \qquad (6)$$

dove $\hat{Y}_d^{k-1}$ e $\hat{Y}_d^{k}$ sono i valori delle stime, riportati nella prima colonna, entro i quali è compresa la stima di interesse $\hat{Y}_d$, ed $\hat{\varepsilon}(\hat{Y}_d^{k-1})$ e $\hat{\varepsilon}(\hat{Y}_d^{k})$ i corrispondenti errori relativi.

È bene precisare che i modelli di interpolazione degli errori sono validi, oltre che per le stime assolute di frequenze e di totali, anche per le stime di frequenze relative e di medie di variabili quantitative riferite all'intera popolazione del dominio di riferimento. Se si vuole calcolare l'errore relativo di una stima riferita a una sottopopolazione differente (ad esempio la popolazione di coloro che presentano una certa modalità di una variabile di interesse, come il titolo di studio) è necessario ricorrere ad un metodo di approssimazione. Infatti, la stima di una frequenza relativa (o di un qualunque

indicatore) riferita al sottogruppo di unità individuate dalla caratteristica *c* (ad esempio le unità in possesso di diploma), è ottenibile come rapporto tra due quantità entrambe stimate:

$$\hat{R}_d = \frac{\hat{N}_d}{\hat{P}_d},$$

in cui $\hat{P}_d$ è la stima del numero di persone che presentano la caratteristica *c* nel dominio d, $\hat{N}_d$ è la stima del numero di persone che presentano una particolare modalità di interesse tra coloro che hanno la caratteristica *c*; $\hat{R}_d$ è l'indicatore definito come rapporto tra $\hat{N}_d$ e $\hat{P}_d$ (per esempio la stima della frequenza relativa dei diplomati occupati).

Una valutazione approssimata[4] dell'errore di $\hat{R}_d$, valida sotto l'ipotesi di incorrelazione tra $\hat{R}_d$ e $\hat{P}_d$, si può ottenere come:

$$\hat{\varepsilon}\left(\hat{R}_d\right) = \sqrt{\hat{\varepsilon}^2\left(\hat{N}_d\right) - \hat{\varepsilon}^2\left(\hat{P}_d\right)},$$

in cui $\hat{\varepsilon}\left(\hat{N}_d\right)$ e $\hat{\varepsilon}\left(\hat{P}_d\right)$ si possono calcolare utilizzando la (5).

Prospetto 1 - Valori dei coefficienti a, b e dell'indice di determinazione R^2 (%) delle funzioni utilizzate per l'interpolazione degli errori campionari delle stime di frequenze per totale Italia, ripartizione geografica, sesso e classi di età

Domini di stima	a	b	R^2
RIPARTIZIONI GEOGRAFICHE			
Nord	6,193417	-0,868940	95,5
Centro	5,739898	-0,825463	94,1
Sud e Isole	6,242355	-0,891710	94,2
SESSO			
Maschi	6,409342	-0,938973	92,3
Femmine	7,840801	-1,044683	95,9
CLASSI DI ETÀ			
0-34	8,404411	-1,102602	92,3
35-54	6,672190	-0,970390	93,7
55 e oltre	7,052031	-0,974595	94,2
TOTALE	**7,092994**	**-0,978538**	**92,7**

Prospetto 2 - Valori interpolati degli errori campionari relativi percentuali delle stime di frequenze per totale Italia, ripartizione geografica, sesso e classi di età

	Domini territoriali				Sesso		Classi di età		
STIME	Italia	Nord	Centro	Sud e Isole	Maschi	Femmine	4-34	35-49	50-67
10.000	38,3	40,5	39,4	37,3	32,6	41,0	41,7	32,2	38,2
20.000	27,3	29,9	29,6	27,4	23,6	28,6	28,4	23,0	27,3
30.000	22,4	25,1	25,0	22,9	19,5	23,1	22,7	18,9	22,4
40.000	19,4	22,2	22,2	20,1	17,0	19,9	19,4	16,4	19,4
50.000	17,4	20,1	20,3	18,2	15,3	17,7	17,2	14,8	17,4
60.000	15,9	18,6	18,8	16,8	14,1	16,1	15,5	13,5	16,0
70.000	14,8	17,4	17,6	15,7	13,1	14,9	14,3	12,5	14,8
80.000	13,8	16,4	16,7	14,8	12,3	13,9	13,2	11,7	13,9
90.000	13,1	15,6	15,9	14,0	11,6	13,0	12,4	11,1	13,1
100.000	12,4	14,9	15,2	13,4	11,1	12,3	11,7	10,5	12,4
200.000	8,8	11,0	11,4	9,8	8,0	8,6	8,0	7,5	8,9
300.000	7,3	9,2	9,7	8,2	6,6	6,9	6,4	6,2	7,3
400.000	6,3	8,1	8,6	7,2	5,8	6,0	5,5	5,4	6,3
500.000	5,6	7,4	7,8	6,5	5,2	5,3	4,8	4,8	5,7

[4] Si veda: P.D. Falorsi, S. Falorsi (1996) 'Indagine sulle forze di lavoro: descrizione della strategia di campionamento e valutazione dell'errore campionario dei principali indicatori provinciali del mercato del lavoro', 1996, ISTAT-Documenti).