Nota metodologica OFP – I Edizione

M. Centra e V. Gualtieri

Obiettivo dell'indagine è di ricostruire le caratteristiche strutturali dell'offerta di formazione professionale regionale.
Per soddisfare tale obiettivo è stata progettata e realizzata un'indagine campionaria sulle strutture di formazione accreditate sul territorio italiano.
La strategia campionaria adottata per la realizzazione dell'indagine sull'offerta di formazione professionale è descritta di seguito: viene innanzitutto definita la popolazione di riferimento e le tecniche utilizzate per la sua identificazione, successivamente sono descritte le principali caratteristiche del piano di campionamento, in conclusione si dettagliano le scelte operate e la tecnica adottata per la determinazione dei pesi di riporto all'universo (fase di stima).

Popolazione di riferimento

La popolazione di riferimento dell'indagine è costituita dalle strutture di formazione professionale, presenti sull'intero territorio nazionale al 31 dicembre 2011, attive e accreditate a livello regionale.
L'archivio relativo alla popolazione di riferimento è stato predisposto dall' Isfol a partire dagli elenchi acquisiti presso le regioni[1].
In dettaglio, la consistenza e la composizione della popolazione di riferimento è stata ricavata, partendo dagli elenchi su menzionati, tramite una rilevazione di screening. Tale rilevazione, effettuata con tecnica CATI, ha verificato l'operatività delle strutture formative nel triennio 2009-2011 quantificando alcune variabili (fatturato e n. allievi) con riferimento all'anno 2011.
Tramite lo screening è stata stimata una popolazione di interesse composta da **3.892** strutture formative accreditate e attive, che costituiscono quindi la popolazione di riferimento dell'indagine.

L'unità di rilevazione è dunque identificata come una struttura di formazione professionale presente sul territorio nazionale che:
- è presente negli archivi in possesso dell'Isfol delle strutture accreditate a livello regionale al 31 dicembre 2011;
- era operativa nel 2012;
- era operativa almeno dal 2011;
- ha organizzato e avviato almeno una attività formativa nel triennio 2009-2011.

Sono state quindi escluse dalla rilevazione le strutture di formazione presenti nel territorio ma non accreditate o accreditate ma non attive.

I dati ricavati dallo screening sono stati utilizzati sia nella fase di disegno del campione sia per costruire il set di informazioni ausiliarie utilizzate nel calcolo dei pesi di riporto all'universo.

[1] Gli elenchi regionali sono stati acquisiti dall'Isfol in momenti differenti e presentano una forte disomogeneità sia nella quantità che nella qualità dei dati contenuti. Per tale ragione, come sarà meglio dettagliato successivamente, si è reso necessario un considerevole lavoro di normalizzazione delle informazioni da utilizzare sia nella fase di disegno sia nella procedura di stima.

Il disegno campionario

La numerosità del campione totale, da intervistare attraverso tecnica CAPI, è stata fissata a **1.200** unità.
Il disegno di campionamento utilizzato è di tipo probabilistico, il piano di campionamento ha previsto un campione stratificato e ha assunto la pianificazione ex-ante dei domini di analisi definendone la numerosità campionaria, vincolata alla numerosità predefinita del campione, in modo da garantire un livello predeterminato di attendibilità delle stime nei domini; sul piano metodologico tale attività si è servita delle opportune tecniche di allocazione negli strati di un campione di numerosità fissata.
La pianificazione dei domini di studio si è dunque concretizzata in un problema di allocazione del campione negli strati, coincidenti con le regioni.
Il problema dell'allocazione è stato risolto ricorrendo a una specifica procedura[2] in grado di garantire l'omogeneità degli errori campionari tra i domini. Nello specifico la procedura determina un'allocazione di compromesso tra l'allocazione uniforme e quella proporzionale, ed ha fatto sì che all'interno di ogni strato vi fossero un numero di centri di formazione tale da garantire un errore campionario approssimativamente costante. Allo stesso tempo la procedura di allocazione ha consentito di controllare l'effetto del disegno del campione non proporzionale[3] .

Tab. 1 Allocazione del campione nelle regioni

Regione	Popolazione	Allocazione complessa	
		Campione	Coefficiente di variazione
Piemonte	454	92	28.0%
Valle D'Aosta	10	9	33.3%
Lombardia	420	81	30.0%
Bolzano	43	31	28.8%
Trento	80	45	29.8%
Veneto	391	84	29.0%
Friuli V. G.	39	30	26.7%
Liguria	61	41	27.1%
Emilia R.	119	59	27.9%
Toscana	442	82	29.9%
Umbria	125	60	28.0%
Marche	154	62	29.5%
Lazio	212	65	31.1%
Abruzzo	68	42	28.8%
Molise	36	25	33.6%
Campania	199	62	31.7%
Puglia	263	70	30.8%
Basilicata	95	50	29.4%
Calabria	120	60	27.5%
Sicilia	479	100	26.7%
Sardegna	82	50	26.7%

n=1200; N=3892; p=0.1; lambda=0.2; var(ccs)=0.000052

Dell'allocazione si è naturalmente tenuto conto anche in fase di costruzione dello stimatore, che ha permesso di riportare il campione alla distribuzione osservata nella popolazione.

[2] Centra M., Falorsi, P.D. (a cura di), Strategie di campionamento per il monitoraggio e la valutazione delle politiche, Roma, Isfol, 2008 (Temi e Strumenti)
[3] La fase di allocazione si serve della procedura contenuta in Centra, Falorsi (2007), pp. 24-32.

Fase di stima

Al campione dei rispondenti è applicato uno stimatore in grado di ricondurre i risultati della rilevazione alla popolazione di riferimento. La costruzione dello stimatore ha previsto il ricorso a tecniche di calibrazione, particolarmente potenti sia per consentire al campione di ricostruire il profilo della popolazione cui è riferito, sia per correggere eventuali fenomeni di distorsione.
La messa a punto della strategia di stima ha previsto quindi l'uso di stimatori indiretti che utilizzano informazioni ausiliarie correlate con le variabili oggetto d'indagine. In particolare, si è fatto riferimento allo stimatore di ponderazione vincolata o calibrato (cfr. Deville and Särndal, 1992).
La struttura generale della procedura è articolata come segue:

1. Determinazione di un peso base definito come l'inverso della probabilità di inclusione di ogni unità campionate;
2. Correzione per mancata risposta totale: procedura che permette di correggere il peso base per gli effetti distorsivi indotti dalle mancate risposte, rispettando così la struttura del campione teorico.
3. Determinazione del peso finale in base alla metodologia degli stimatori calibrati. Tale metodologia, basata sugli stimatori parzialmente assistiti da modello, sulla base degli stimatori di regressione[4], consente di vincolare il campione sia alla struttura della popolazione di riferimento utilizzata nella fase di stratificazione che a strutture derivate da fonti esterne, e non necessariamente considerate nel disegno.

L'approccio predittivo permette la messa a punto di stimatori calibrati basati su una serie di informazioni ausiliarie; oltre a sfruttare le informazioni delle variabili ausiliare riducendo la varianza campionaria, tale classe di stimatori gode di una serie di proprietà tra le quali quella della calibrazione, secondo la quale le stime dei totali delle variabili ausiliarie utilizzate come regressori, corrispondono ai totali noti. In tal modo è possibile calibrare la popolazione stimata rispetto ai totali noti ricavati dalla popolazione di riferimento, disaggregati secondo specifiche caratteristiche. Gli aggregati di riferimento, utilizzati come totali noti dalla procedura di calibrazione, sono stati ricavati dalla fase di screening. Le informazioni ausiliarie utilizzate nella costruzione dello stimatore calibrato sono riportate nello schema seguente:

Schema 1 Informazione ausiliarie per la procedura di calibrazione

Descrizione	**Modalità**
REGIONE	21 regioni e provincie autonome
ALLIEVI	Numero
FILIERE FORMATIVE	Obbligo formativo, formazione superiore, formazione continua, orientamento, soggetti svantaggiati

Il piano di vincoli impone che all'interno di ciascuna regione il campione riproduca la distribuzione osservata nella popolazione secondo la filiera formativa e il numero totale di allievi formati nel 2011.
Lo stimatore così ottenuto, applicato come coefficiente moltiplicativo delle unità campionarie, ha permesso di produrre stime sulla popolazione di riferimento in modo che gli aggregati riferiti a ciascuna nidificazione riportata nel piano di calibrazione, coincidessero con i corrispondenti totali noti ricavati dalla fase di screening.

[4] Dorfman A.H., Royall R.M., Valliant R., Finite Population Sampling and Inference: a Prediction Approach, New York , John Wiley & Sons, 2000

Valutazione della affidabilità delle stime

Come ogni indagine campionaria, le stime fornite sono soggette a errore di campionamento. La procedura per il calcolo dell'errore campionario associato alle stime prodotte è fondata sulle usuali tecniche note in letteratura derivanti dalla scelta dello stimatore proposto. Nello specifico, la proprietà cardine degli stimatori calibrati è la convergenza asintotica allo stimatore di regressione generalizzato. Grazie a tale proprietà, per campioni di importanti dimensioni, è possibile utilizzare tutti i risultati analitici noti per lo stimatore di regressione generalizzata, tra i quali vi è la forma analitica della varianza dello stimatore di regressione generalizzata che può essere utilizzata per calcolare l'errore delle stime prodotte dallo stimatore di ponderazione vincolata (Deville, Särndal, 1992).
Il livello dell'attendibilità delle stime è misurato tramite il coefficiente di variazione, *CV(p)* riferito ad una generica stima di una frequenza relativa p nella popolazione. E' stata calcolata una serie sufficientemente numerosa di coppie (p; CV) per ciascuno dei domini di studio considerati nell'analisi: ripartizione geografica, classe di allievi, filiera formativa, oltre al dominio totale coincidente con l'intera popolazione. Le serie (p; CV) sono state interpolate tramite una procedura di regressione robusta secondo un modello logaritmico-lineare:

$$\ln (CV)^2 = b_0 + b_1 ln(p) \qquad (1)$$

tramite il quale sono stati stimati i parametri b_0 e b_1, per ciascun dominio, come riportato nella tabella seguente:

Tab. 1 Coefficienti stimati per il calcolo del CV

Dominio		b_0	b_1
Totale		-8,25	-1,46
Area	Nord ovest	-6,06	-1,21
	Nord est	-6,26	-1,22
	Centro	-6,63	-1,23
	Mezzogiorno	-7,32	-1,41
Allievi	Non disponibile	-4,64	-1,16
	Fino a 50	-5,99	-1,14
	Tra 51 e 100	-5,77	-1,25
	Tra 101 e 250	-6,46	-1,29
	Tra 251 e 500	-6,21	-1,28
	Tra 501 e 1000	-6,02	-1,23
	Oltre 1000	-6,34	-1,30
Filiera	Obbligo formativo	-7,22	-1,30
	Formazione Superiore	-7,60	-1,16
	Formazione Continua	-8,19	-1,46
	Orientamento	-5,77	-1,14
	Svantaggiati	-4,94	-1,35

In tal modo è possibile calcolare il CV di una generica stima p riferita ad uno specifico dominio (nord-ovest, nord-est, ecc.) applicando la funzione (1) nella forma:

$$CV(p) = \sqrt{\exp\left(b_0 + b_1 ln(p)\right)}$$

applicando i valori dei parametri b_0 e b_1 corrispondenti al dominio considerato.
Ad esempio, il CV di una frequenza relativa stimata pari a 0,13, riferita ai centri formativi che hanno attivato la filiera dell'orientamento è pari a:

$$CV(p = 0{,}13) = \sqrt{\exp\left(-5{,}77 - 1{,}14 \cdot ln(p)\right)} = 0{,}18$$

Tramite il CV così ottenuto è possibile calcolare agevolmente l'intervallo di confidenza al 95% ($z_\alpha = 1{,}96$):

$$p - CV(p) \cdot p \cdot z_\alpha \leq p \leq p + CV(p) \cdot p \cdot z_\alpha$$