

Esempio di campionamento complesso

(Appunti Metodologia: lezione 6)

1. Introduzione

Il campionamento riguarda un'indagine sulla mobilità in provincia di Bologna (1996).

La popolazione di riferimento sono le 375.729 famiglie residenti in provincia di Bologna, corrispondenti a 906.018 individui (dati 1994).

La ricerca consiste nella somministrazione tramite intervistatori di un questionario familiare, di un questionario individuale per ciascuno dei membri della famiglia e di un diario degli spostamenti giornalieri, anch'esso per ciascun componente della famiglia. La somministrazione dei questionari avviene come segue: l'intervistatore si reca una prima volta presso la famiglia e in questa occasione compila il questionario familiare e i questionari individuali dei membri presenti nell'abitazione. Lascia inoltre i diari degli spostamenti – uno per ciascun componente – che saranno autocompilati dai soggetti con riferimento al giorno successivo. In questa occasione, l'intervistatore fissa un appuntamento per il ritiro dei diari e per la somministrazione degli eventuali questionari individuali rimanenti. L'intervista ad una famiglia si considera conclusa solo quando si sono ottenute tutte le interviste e i diari individuali. Le informazioni richieste nei questionari riguardano gli stili di vita delle famiglie, i modelli di consumo, le spese per mobilità e i comportamenti di mobilità (mezzi, tempi, numerosità degli spostamenti, destinazioni, ecc.)

Il campione è composto da 3.500 famiglie, pari al 9,3% dell'universo di riferimento.

2. Considerazioni preliminari circa la formazione del campione

Per effettuare un campionamento casuale sarebbe stato necessario costruire la lista delle famiglie residenti. L'unico modo per costruire questa lista sarebbe stato quello di chiedere a ciascuno dei 60 comuni presenti in provincia la lista delle famiglie residenti in modo da costruire una lista complessiva. Ragioni di tempi e di costi, oltre agli ostacoli frapposti dalla legge sulla privacy, hanno escluso di poter percorrere questa strada.

L'unica alternativa percorribile è stato il ricorso ad un campionamento sistematico, che per la sua semplicità e praticità poteva essere effettuato anche da un impiegato comunale o, per via informatica, presso i singoli comuni. Era sufficiente determinare la quota di famiglie da intervistare in ciascun comune.

Anche la seconda alternativa presentava alcuni limiti. Occorre infatti ottenere la collaborazione di tutti i 60 sindaci. Va tenuto presente che l'indagine veniva presentata alle famiglie campione proprio da una lettera del sindaco. E' evidente che sarebbe stato sufficiente che uno o due sindaci frapponessero difficoltà per ritardare anche di molto i preliminari dell'indagine.

Si è ritenuto pertanto di ridurre il numero dei comuni interessati, tenendo anche presente che si pensava di replicare negli anni successivi l'indagine. In tal caso i comuni sarebbero stati ruotati, cioè alcuni comuni oggetto d'indagine sarebbero stati sostituiti da comuni esclusi in questa ricerca.

3. Determinazione dei comuni oggetto d'indagine.

Si deve innanzitutto considerare che uno degli scopi dell'indagine era quello di ottenere informazioni in vista della formulazione dei piani traffico richiesti da una normativa nazionale ai comuni con più di 30000 abitanti e comunque ai comuni confinanti con i capoluoghi.

Perciò questi comuni dovevano essere compresi nel campione. Si tratta di 18 comuni.

I restanti comuni sono tra loro molto diversi per la loro collocazione geografica che a sua volta influisce certamente sui comportamenti di mobilità.

Si è deciso pertanto di dividere questi comuni in comuni di pianura (16), in comuni di collina (13), comuni di montagna (13).

Ciascuna di queste tre categorie doveva essere rappresentata nel campione finale. Si è deciso di selezionare 2 comuni per ciascuna delle categorie.
In totale i comuni considerati sono stati 24.

4. Caratteristiche generali del campione

Si è ritenuto che il campione finale di famiglie dovesse essere autoponderante (ogni famiglia della popolazione doveva avere la stessa probabilità di estrazione), così da evitare la determinazione e l'uso a posteriori di pesi campionari.

Si è ritenuto inoltre che il campione dovesse avere numerosità fissa a priori.

5. Disegno di campionamento

I 60 comuni della provincia di Bologna sono suddivisi nelle quattro seguenti aree:

A. Comuni autorappresentativi, costituiti dai seguenti 18 comuni: Bologna, Imola, Anzola dell'Emilia, Budrio, Calderara, Casalecchio di Reno, Castelmaggiore, Castel San Pietro Terme, Castenaso, Crevalcore, Medicina, Molinella, Ozzano dell'Emilia, Pianoro, San Giovanni in Persiceto, San Lazzaro di Savena, Sasso Marconi, Zola Predosa.

B. Restanti comuni della pianura (16): Argelato, Baricella, Bentivoglio, Castel Guelfo di Bologna, Castello d'Argile, Crespellano, Galliera, Granarolo dell'Emilia, Malalbergo, Minerbio, Mordano, Pieve di Cento, Sala Bolognese, San Giorgio di Piano, San Pietro in Casale, Sant'Agata Bolognese.

C. Restanti comuni della collina (13): Bazzano, Borgo Tossignano, Casalfiumanese, Castel del Rio, Castello di Serravalle, Dozza, Fontanelice, Loiano, Marzabotto, Monte San Pietro, Montereenzio, Monteveglio, Savigno.

D. Restanti comuni della montagna (13): Camugnano, Castel d'Aiano, Castel di Casio, Castiglione dei Pepoli, Gaggio Montano, Granaglione, Grizzana, Lizzano in Belvedere, Monghidoro, Monzuno, Porretta Terme, San Benedetto Val di Sambro, Vergato.

A ciascuna delle quattro aree è stata assegnata una quota campionaria proporzionale al numero complessivo di famiglie residenti nell'area.

Per l'area A, tale quota campionaria è stata ripartita tra i diversi comuni proporzionalmente al numero di famiglie residenti. Si è proceduto successivamente ad estrarre in modo sistematico dalle anagrafi di ciascun comune le famiglie campione. A questo scopo è stato individuato un passo di campionamento costante in ciascun comune (50). Tale passo di campionamento assicura l'estrazione di un numero di famiglie pari a circa 2,1 volte la quota campionaria come sopra determinata. Ciò consente la formazione di un campione di riserva cui si ricorre quando e se si rendono necessarie sostituzioni nel corso dell'indagine. L'assegnazione delle famiglie al campione effettivo o alla riserva avviene con estrazione casuale.

Per le aree B, C, D, si è proceduto ad un campionamento su due stadi con procedura PPS (Probability Proportional to Size).

Al primo stadio in ciascuna delle aree sono stati estratti due comuni con metodo casuale semplice e probabilità proporzionale alle dimensioni dei comuni (numero di famiglie residenti)

Nelle tre aree (B, C, D) le quote campionarie come sopra determinate sono state attribuite in parti uguali ai due comuni estratti, così da assicurare che le famiglie appartenenti alle tre coppie di comuni avessero una probabilità d'estrazione inversamente proporzionale alla dimensione del comune.

Si è proceduto poi come per l'area A selezionando le famiglie campione con metodo sistematico dalle anagrafi comunali. Anche qui è stato individuato per ciascun comune un passo di campionamento tale da assicurare un extracampione del 100% circa. L'attribuzione delle famiglie al campione effettivo o alla riserva avviene con le stesse modalità già esposte per l'area A.

In conclusione si trattava di un campione a due stadi, stratificato nel primo stadio (21 strati, di cui 18 di ampiezza $n=1$), con ricorso al campionamento casuale semplice a probabilità variabile all'interno dei singoli strati, e unità di campionamento i comuni; sistematico nel secondo stadio, con unità di campionamento le singole famiglie.

Il campione finale così formato è autoponderante e perciò non richiede il ricorso a pesi campionari per riportare i risultati all'insieme della provincia.

Nella Tab. 1 allegata sono evidenziati i comuni della provincia interessati dall'indagine. Per ciascuno è indicata la popolazione residente e il numero di famiglie residenti al 31.12.1994 (dati forniti dal Servizio Informativo, statistica e relazioni con il pubblico della Regione Emilia Romagna). Sono riportati inoltre tutti i dati utilizzati e da utilizzare per portare a compimento il campionamento.

6. Come sono calcolati i dati di tab. 1.

- **Dati di colonna C:** dapprima si individuano le quote campionarie per le quattro aree: comuni autorappresentativi, comuni di pianura, comuni di collina, comuni di montagna. Le quote si individuano dividendo il numero di famiglie nello strato (colonna B) per il numero di famiglie dell'intera provincia e moltiplicando il risultato per il totale famiglie del campione (prefissato a 3500). Tra i comuni autorappresentativi questi valori si ottengono dividendo il numero di famiglie del singolo comune per il totale famiglie di questo strato e moltiplicando il risultato per la quota di famiglie di questo strato (determinato nel passo precedente. Tra i sei comuni estratti per le tre aree (pianura, collina, montagna) le quote campionarie vengono determinate dividendo in parti uguali le quote determinate al primo passo per le tre aree (330, 186, 192 rispettivamente) tra i due comuni estratti per ciascuna area.

- **Dati di colonna F:** Viene determinato un passo di campionamento che porti all'estrazione di un campione doppio di quello preventivato. La scelta del valore 50 piuttosto che 53 è ovviamente dettata dalla maggior facilità d'uso di un passo multiplo di 10, anche se con questo passo il campione risulta 2,1 volte quello richiesto. Il passo per i sei comuni delle tre aree è determinato in modo da ottenere un campione di circa 2,1 volte quello effettivo, per uniformità con quanto fatto tra i comuni autorappresentativi. I dati dei quattro strati iniziali si ottengono per somma dei dati dei singoli comuni degli strati.

- **Dati di colonna E:** si ottengono dividendo i dati di colonna B per quelli di colonna F.

- **Dati di colonna D:** si ottengono per differenza tra quelli di colonna E e quelli di colonna C.

- **Dati di colonna G:** si tratta della famiglia iniziale da cui si inizia ad applicare il passo di campionamento di colonna F. Se le famiglie sono in ordine alfabetico, un passo di 50 significa che si può iniziare da una qualsiasi delle prime 50 famiglie. Il numero è stato determinato con estrazione casuale nell'intervallo determinato dal passo di campionamento.

Infine in **colonna H** è riportata la probabilità di inclusione nel campione delle famiglie appartenenti a ciascuno degli strati e dei singoli comuni. Per i sei comuni delle tre aree residue queste probabilità verranno fornite nella successiva tab. 2.

7. Significato dei dati di tab. 2.

Nelle aree restanti (pianura collina, montagna), vengono estratti, come detto, 2 comuni per ciascuna area.

Si è detto inoltre che l'estrazione avviene a probabilità variabili. Per questo tipo di estrazione si procede così. Esemplichiamo per la pianura. Invece di mettere in un'urna i nomi dei 16 comuni di quest'area, per ciascuno dei 16 comuni si introducono nell'ipotetica urna tanti numeri quante sono le famiglie di quel comune. Il primo comune ha 2872 famiglie e a questo comune vengono associati i numeri da 1 a 2872. Il secondo comune ha 2082 famiglie e gli si assegneranno 2082 numeri, quelli compresi tra 2873 e 4954. si procede così fino all'ultimo comune. Al termine l'ipotetica urna conterrà i numeri compresi tra 1 e 35547, tanti cioè quante sono le famiglie di quest'area. Questi numeri sono associati all'uno o all'altro comune. Sarà sufficiente a questo punto estrarre casualmente due numeri nell'intervallo 1-35547 e vedere a quali comuni sono associati. I due comuni associati sono i comuni da estrarre. (Naturalmente se il secondo numero appartiene allo stesso comune del primo numero estratto, questo secondo numero verrà scartato e si procederà ad una ulteriore estrazione, fino a che si siano estratti due comuni diversi. E' evidente che i comuni più grandi hanno più numeri associati nell'urna, quindi una probabilità di estrazione più elevata rispetto ad un comune di dimensioni inferiori.

Nelle colonne "probabilità" sono riportate le probabilità di estrazione al primo stadio, al secondo stadio (quando si estrarranno le quote di famiglie determinate in Tab. 1) e la probabilità d'inclusione per le unità di secondo stadio (le famiglie). Come si vede, al primo stadio le probabilità di selezione sono proporzionali al numero di famiglie del campione. Al secondo stadio invece, le probabilità di selezione sono inversamente proporzionali al numero di famiglie (va notato che per ciascun comune questa probabilità si calcola dividendo 165 casi - tanti se ne devono estrarre secondo tab.1 da ciascun comune - per il numero delle famiglie e moltiplicando per 2 - tanti sono infatti i comuni che vengono estratti dall'area).

La probabilità di inclusione delle famiglie di quest'area è sempre del 9,3%, qualsiasi comune venga estratto, per effetto dell'andamento divergente delle probabilità di selezione al primo e secondo stadio. Questa probabilità di inclusione coincide con quella dei comuni autorappresentativi.

Tab. 1 - Comuni della provincia di Bologna facenti parte del campione, numerosità campionarie per comune (effettivi e sostituzioni), passo di campionamento per comune, primo caso da estrarre

	A	B	C	D	E	F	G	H
	Popolazione residente	Famiglie residenti	Campione effettivo	Campione sostituzioni	Campione da estrarre			(C/B)
Totale provincia	906018	375729	3500	3966	7466			0,0093
di cui:								
Comuni autorappresentativi	713875	299634	2792	3192	5984			0,0093
Strato comuni della pianura	94042	35547	330	334	664			0,0093
Strato comuni della collina	50439	19955	186	225	411			0,0093
Strato comuni della montagna	47662	20593	192	215	407			0,0093
								Probabilità inclusione
Comuni autorappresentativi				(E-C)	(B/F)	Passo di estrazione	Primo campio caso ne	
Anzola dell'Emilia	10013	3719	35	39	74	50	21	0,0094
Bologna	390434	174732	1628	1866	3494	50	4	0,0093
Budrio	14785	5630	52	60	112	50	41	0,0092
Calderara di Reno	11379	4306	40	46	86	50	32	0,0093
Casalecchio di Reno	33423	13560	126	145	271	50	46	0,0093
Castel Maggiore	15242	5918	55	63	118	50	14	0,0093
Castel San Pietro Terme	18604	6874	64	73	137	50	45	0,0093
Castenaso	13506	4899	46	51	97	50	47	0,0094
Crevalcore	11518	4388	41	46	87	50	50	0,0093
Imola	63614	24783	231	264	495	50	9	0,0093
Medicina	12706	4812	45	51	96	50	28	0,0094
Molinella	12331	5014	47	53	100	50	11	0,0094
Ozzano dell'Emilia	9911	3851	36	41	77	50	31	0,0093
Pianoro	14853	5704	53	61	114	50	23	0,0093
San Giovanni in Persiceto	22983	8697	81	92	173	50	21	0,0093
San Lazzaro di Savena	29283	11386	106	121	227	50	45	0,0093
Sasso Marconi	13242	5114	48	54	102	50	4	0,0094
Zola Predosa	16048	6247	58	66	124	50	29	0,0093
Totale autorappresentativi	713875	299634	2792	3192	5984			
Comuni della pianura estratti								
Galliera	4728	1988	165	166	331	6	3	
Sant'Agata Bolognese	5258	2001	165	168	333	6	1	
Totale pianura	94042	35547	330	334	664			
Comuni della collina estratti								
Monte San Pietro	8631	3327	93	102	195	17	4	
Savigno	2372	1082	93	123	216	5	5	
Totale collina	50439	19955	186	225	411			
Comuni della montagna estratti								
Lizzano in Belvedere	2314	1062	96	116	212	5	1	
San Benedetto Val di Sambro	4264	1757	96	99	195	9	7	
Totale montagna	47662	20593	192	215	407			

Tab. 2 - Assegnazione di numeri casuali ai comuni della pianura, collina e montagna (esclusi quelli ricompresi tra i comuni autorappresentativi)

Comuni della pianura

	Famiglie residenti	Numeri casuali		Probabilità		Probabilità
		da	a	1° stadio	2° stadio	Inclusione
Argelato	2872	1	2872	0,0808	0,1149	0,0093
Baricella	2082	2873	4954	0,0586	0,1585	0,0093
Bentivoglio	1154	4955	6108	0,0325	0,2860	0,0093
Castel Guelfo	1052	6109	7160	0,0296	0,3137	0,0093
Castello d'Argile	1551	7161	8711	0,0436	0,2128	0,0093
Crespellano	2654	8712	11365	0,0747	0,1243	0,0093
Galliera	1988	11366	13353	0,0559	0,1660	0,0093
Granarolo dell'Emilia	3170	13354	16523	0,0892	0,1041	0,0093
Malalbergo	2629	16524	19152	0,0740	0,1255	0,0093
Minerbio	2779	19153	21931	0,0782	0,1187	0,0093
Mordano	1449	21932	23380	0,0408	0,2277	0,0093
Pieve di Cento	2453	23381	25833	0,0690	0,1345	0,0093
Sala Bolognese	1908	25834	27741	0,0537	0,1730	0,0093
San Giorgio di Piano	2073	27742	29814	0,0583	0,1592	0,0093
San Pietro in Casale	3732	29815	33546	0,1050	0,0884	0,0093
Sant'Agata Bolognese	2001	33547	35547	0,0563	0,1649	0,0093

Comuni della collina

	Famiglie residenti	Numeri casuali		Probabilità		Probabilità
		da	a	1° stadio	2° stadio	Inclusione
Bazzano	2160	1	2160	0,1082	0,0861	0,0093
Borgo Tossignano	1065	2161	3225	0,0534	0,1746	0,0093
Casalfiumanese	1006	3226	4231	0,0504	0,1849	0,0093
Castel del Rio	479	4232	4710	0,0240	0,3883	0,0093
Castello di Serravalle	1317	4711	6027	0,0660	0,1412	0,0093
Dozza	1867	6028	7894	0,0936	0,0996	0,0093
Fontanelice	653	7895	8547	0,0327	0,2848	0,0093
Loiano	1411	8548	9958	0,0707	0,1318	0,0093
Marzabotto	2278	9959	12236	0,1142	0,0817	0,0093
Monte San Pietro	3327	12237	15563	0,1667	0,0559	0,0093
Monterenzio	1736	15564	17299	0,0870	0,1071	0,0093
Montevoglio	1574	17300	18873	0,0789	0,1182	0,0093
Savigno	1082	18874	19955	0,0542	0,1719	0,0093

Comuni della montagna

	Famiglie residenti	Numeri casuali		Probabilità		Probabilità
		da	a	1° stadio	2° stadio	Inclusione
Camugnano	941	1	941	0,0457	0,2040	0,0093
Castel d'Aiano	815	942	1756	0,0396	0,2356	0,0093
Castel di Casio	1269	1757	3025	0,0616	0,1513	0,0093
Castiglione dei Pepoli	2368	3026	5393	0,1150	0,0811	0,0093
Gaggio Montano	1897	5394	7290	0,0921	0,1012	0,0093
Granaglione	1008	7291	8298	0,0489	0,1905	0,0093
Grizzana	1375	8299	9673	0,0668	0,1396	0,0093
Lizzano in Belvedere	1062	9674	10735	0,0516	0,1808	0,0093
Monghidoro	1360	10736	12095	0,0660	0,1412	0,0093
Monzuno	1852	12096	13947	0,0899	0,1037	0,0093
Porretta Terme	2343	13948	16290	0,1138	0,0819	0,0093
San Benedetto Val di Sambro	1757	16291	18047	0,0853	0,1093	0,0093
Vergato	2546	18048	20593	0,1236	0,0754	0,0093