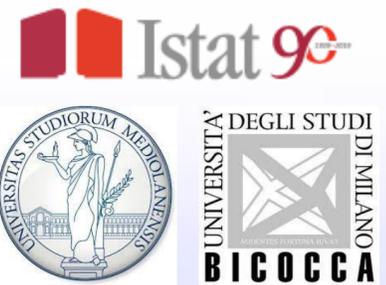


La stima nei domini di studio territoriali: dall'allocazione di compromesso all'IPNLP



P. M. Chiodini¹ - G. Manzi² - B. M. Martelli³ - F. Verrecchia⁴

¹ Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca |

² Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano |

³ Già Istat | ⁴ Ufficio Territoriale per il Piemonte e la Valle d' Aosta, la Lombardia e la Liguria (Nord Ovest), Istat

Introduzione

Il disegno di campionamento dell'indagine sul clima di fiducia delle imprese (BCS) è stato recentemente rivisto con particolare riferimento ai metodi di allocazione campionaria. È stata adottata una allocazione di compromesso, l'Allocazione robusta con soglia di strato uniforme [2] (ROAUST – acronimo in inglese) che nel poster viene confrontata con altri metodi attraverso simulazione Montecarlo, utilizzando dati di stratificazione provenienti dal registro ASIA.

1. Il frame per la simulazione

ASIA fornisce una lista completa delle imprese italiane con più di 10 addetti. Il frame risultante comprende 90mila imprese (circa il 18% delle imprese e il 77% degli addetti). È un frame ideale per le indagini BCS europee in quanto fornisce indicazioni per le tre variabili di stratificazione: (i) dimensione in termini di addetti (3 classi); (ii) ripartizione territoriale (4 classi); (iii) settori NACE (19 classi). In questo modo si hanno 228 strati dei quali due risultano vuoti. Nella simulazione si producono campioni con numerosità predeterminata ($n = 4.000$) e la variabile numero di addetti viene usata come proxy del PIL e quindi del clima di fiducia delle imprese.

2. Metodi di allocazione

Oltre ai tradizionali semplici metodi di allocazione (uniforme, proporzionale), a quelli generalmente considerati più performanti (Neyman) e ad alcuni più specificatamente utilizzati in indagini BCS (il metodo "bounded" ISAE e il Neyman-Bethel), si considerano alcuni metodi utilizzati in indagini sulle imprese in altri paesi come il metodo di Costa [4] (PAUST), quello di Bankier [1] e il metodo NLP (Non Linear Programming) [3]. Questi metodi vengono confrontati con il ROAUST che può essere sintetizzato per campioni di n elementi:

- Si utilizza l'allocazione uniforme per $n_1 < n$ unità campionarie con $n_1 = \alpha n$, $\alpha \in [0, 1]$. La numerosità per lo strato b è quindi pari a: $n_{1h} = \frac{n_1}{H}$ (H è il numero degli strati).
- Si utilizza l'allocazione di Neyman per le restanti unità campionarie.

Riferimenti bibliografici

- Bankier, M.D. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *American Statistician*, 42(3), 174-177.
- Chiodini, P.M., Manzi, G., Verrecchia, F. (2008). Allocazione ottimale robusta con soglia uniforme di strato. *ESeC W.P. N. 5/2008*.
- Choudhry, G.H., Rao, J.N.K., Hidiroglou, M.A. (2012). On Sample Allocation for Efficient Domain Estimation. *Survey Methodology*, 38, 23-29.
- Costa, A., Satorra, A., Ventura, E. (2004). Using Composite Estimators to Improve both Domain and Total Area Estimation. *SORT*, 28, 69-86.
- Ohlsson, E. (1995) Co-ordination of samples using permanent random numbers. In *Business Survey Methods* (eds B. Cox, D. Binder, B. Chinnappa, A. Christianson, M. Colledge and P. Kott), pp. 153-169. New York: Wiley.

3. Piano di simulazione

Il piano di simulazione è strutturato in modo tale da separare i processi di selezione, allocazione e d' inferenza. Questo processo può essere ricondotto alla tecnica *Permanent Random Number* (PRN) proposta da Ohlsson [5] e può essere riassunto nel modo seguente.

Sia P una popolazione suddivisa in H strati, N_1, N_2, \dots, N_H le numerosità degli strati, n_1, n_2, \dots, n_H le numerosità campionarie degli strati. Il piano di simulazione funziona nel seguente modo:

- Si replicano 1000 volte la popolazione P e si indicano con $P_1, \dots, P_p, \dots, P_{1000}$ le repliche;
- Per ogni P_r :
 - Si assegnano un numero PRN distinto per ogni unità delle popolazioni;
 - Si ordinano le unità in ogni strato in senso ascendente rispetto al numero PRN;
 - Si selezionano le unità campionarie in ogni strato in base al metodo di allocazione. Le numerosità campionarie devono essere pari a ≈ 4000 ;
 - Si ottengono le stime \bar{y}_r del numero medio degli addetti;
 - Si ottengono le stime di strato \bar{y}_{rh} del numero medio degli addetti;
- Dalle 1000 stime \bar{y}_r si calcola la radice dell'errore quadratico medio relativo (RRMSE – acronimo in inglese);
- Dai 226 strati si calcolano i 226 RRMSE di strato.

Sulla base dei valori RRMSE di popolazione e di strato si confrontano i diversi metodi di allocazione.

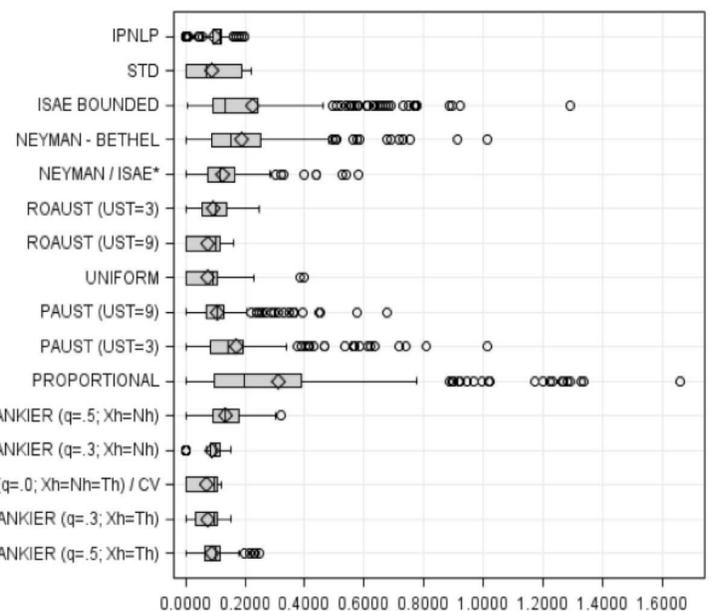
4. Risultati di strato della simulazione

Prendendo in considerazione solo i metodi con un RRMSE di popolazione minore o uguale a 0.70% e considerando il massimo degli RRMSE di strato per il ROAUST($u=9$) come valore di riferimento, il metodo Neyman riporta un valore dell'RRMSE 3 volte più grande, il Bankier un valore più grande della metà e l'IPNLP un valore più grande di un quarto.

In particolare, osservando i metodi IPNLP e il ROAUST($u=9$) risultano differenze già emergenti dal *box and whiskers plot*:

- L'errore per l'IPNLP è centrato e concentrato sul 12%, mentre la maggior parte della distribuzione dell'errore di strato per il ROAUST($u=9$) è concentrata sotto il 12%.
 - il ROAUST($u=9$) è meglio dell'IPNLP in 49% degli strati, risulta uguale nel 23% degli strati e peggiore solo nel 28% degli strati.
- Risultati simili in termini di peso di strato.

	Missing strata #	Max of strata relative Biases	Max of strata RRMSEs	Overall RRMSEs
IPNLP($u=1$)	0	0.0109	0.1978	0.0069
STD = BANKIER ($q=1; X_h = Y_h$)	4	0.0167	0.2181	0.0144
ISAE BOUNDED	6	0.0413	1.2881	0.0250
NEYMAN - BETHEL	0	0.0352	1.0115	0.0264
NEYMAN / ISAE * = ROAUST($u=9$) = BANKIER ($q=1; X_h = Y_h$)	11	0.0155	0.5809	0.0063
ROAUST($u=9$)	0	0.0155	0.2495	0.0064
ROAUST($u=9$)	0	0.0115	0.1588	0.0069
UNIFORM = ROAUST($u=21$) = PAUST($u=21$)	0	0.0112	0.3966	0.0179
PAUST($u=9$)	0	0.0292	0.6743	0.0248
PAUST($u=3$)	0	0.0362	1.0105	0.0355
PROPORTIONAL($u=1$) = PAUST($u=1$)	0	0.0550	1.6587	0.0520
BANKIER ($q=5; X_h = N_h$)	5	0.0155	0.3184	0.0130
BANKIER ($q=3; X_h = N_h$)	5	0.0112	0.1520	0.0099
BANKIER ($q=0; X_h = N_h = T_h$) = CV	5	0.0103	0.1187	0.0098
BANKIER ($q=3; X_h = T_h$)	5	0.0103	0.1526	0.0079
BANKIER ($q=5; X_h = T_h$)	5	0.0116	0.2495	0.0070



RRMSE	Strata %	Persons employed %
IPNLP-ROAUST9 > 0	49%	38%
IPNLP-ROAUST9 = 0	23%	24%
IPNLP-ROAUST9 < 0	28%	38%
Total	100%	100%

RRMSE	IPNLP	ROAUST($u=9$)
Max	0.1978	0.1588
Q3	0.1178	0.1150
Med	0.1144	0.0996
Q1	0.0931	0.0000
Min	0.0000	0.0000
Range	0.1978	0.1588

5. Discussione

Il poster mette in luce come sia l'allocazione di compromesso (es. ROAUST) sia l'allocazione *Interior Point Non Linear Programming* siano di grande interesse per indagini campionarie relative a sub-domini di studio (es. territoriali). L'allocazione proporzionale (e la famiglia PAUST) non sono invece performanti anche in relazione alla correlazione negativa tra numerosità e variabilità di strato, situazione tipica delle indagini sulle imprese che in genere considerano la dimensione aziendale tra le variabili di stratificazione. L'allocazione uniforme, anche per questo motivo, da risultati migliori dell'allocazione proporzionale.

RRMSE (226 strati, 1000 repliche). I risultati presentati risultano sotto referaggio per la pubblicazione su rivista.