

Optimal Sample Selection in the Italian Retail Trade Monthly Survey

*Selezione ottimale del campione nell'indagine mensile
sulle vendite al dettaglio*

Roberto Gismondi

Istat, via Tuscolana 1788, 00177 Roma, Italy, gismondi@istat.it

Anna Rita Giorgi, Tiziana Pichiorri

Istat, viale Liegi 13, 00198 Roma, Italy, angiorgi@istat.it, pichiorri@istat.it

Riassunto: La stima del fatturato mensile delle imprese operanti nel comparto del commercio al dettaglio si basa, attualmente, su un disegno campionario casuale stratificato. Si propone il confronto con strategie alternative, basate su diversi modelli di superpopolazione e la minimizzazione della *Mqe*, con un'applicazione ai dati 2004-2006.

Keywords: Mean squared error, Model, Predictor, Retail trade, Sample

1. Comparison among estimation strategies in retail trade

The retail trade sample survey is carried out by ISTAT since 1996: through a stratified random sampling, in each stratum estimates of monthly turnover indexes are based on the sample mean. In order to compare alternative estimation strategies under a model based approach, a database including 324 enterprises operating in the *Food and beverages* domain - always respondent in the 34 months from January 2004 to October 2006 - has been built, using 2004 and 2005 data for estimating 2006 monthly turnover. Given a population P with size N , the target is the estimation of the population mean $\bar{y}_P$ through a sample s with size n . A general well known regression model is defined as: $y_i = \beta x_i + \varepsilon_i$, with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2 v_i$, $Cov(\varepsilon_i, \varepsilon_j) = c_{ij}$ for each (i) or $(i \neq j)$, where x and v are variables to be specified, with β and σ^2 unknown parameters. If $\bar{y}_s$ is the sample mean, a general linear predictor of the unknown mean is given by:

$$T = N^{-1} [n \bar{y}_s + (N - n) \sum_s a_i y_i] = N^{-1} [n \bar{y}_s + (N - n) U_s] \quad (1)$$

where U_s is a predictor of the not observed mean $\bar{y}_s$ and weights a_i should guarantee unbiasedness: $E(T - \bar{y}_P) = 0$ and the lowest $Mse(T) = E(T - \bar{y}_P)^2$. Possible solutions depend on hypotheses on c_{ij} and v_i . Since Mse is evaluated according to the model, for some n its level can depend on the particular selected sample (*purposive* sampling). If one supposes $c_{ij} = 0$ for each $(i \neq j)$, it is well known (Cicchitelli *et al.*, 1992) that $U_s^* = \bar{x}_s \beta^*$, where the BLUP of β and the Mse of the optimal T^* will be, respectively:

$$\beta^* = (\sum_s x_i y_i / v_i) (\sum_s x_i^2 / v_i)^{-1} \quad Mse(T^*) = N^{-2} [x_s^2 (\sum_s x_i^2 / v_i)^{-1} + v_s] \sigma^2 \quad (2)$$

where $x_{\bar{s}}$ and $v_{\bar{s}}$ are sums over units not in the sample. The recourse to the actual A) sample mean is coherent with $x_i=v_i=1$ for each i . When $v=1$, $v=x$ or $v=x^2$ one gets, respectively, that β is estimated by a B) regression through the origin, C) a ratio estimator and a D) mean of ratios y/x , all based on sample units only. In the first 2 cases the lowest Mse is got choosing the n x -largest units, in the third one it should maximise $2N\bar{x}_s\bar{x}_P+(n-1)S_{sx}^2$, where S_{sx}^2 is the sample x -variance. The further *generalised regression estimator* E) $\bar{y}_s + \beta^*(\bar{x}_P - \bar{x}_s)$ - asymptotically unbiased under whatever sampling design - has been compared according to a (SRS) design in each stratum. More generally, if in the model $c_{ij} \neq 0$, the analytical form of the optimal predictor cannot be easily achieved. Putting in (1) $a_i=a$ for each i , we have (Gismondi, 2007) that:

$$Mse(T) = \bar{x}_{\bar{s}}^2 \bar{x}_s^{-2} Var(\bar{y}_s) + Var(\bar{y}_{\bar{s}}) - 2\bar{x}_{\bar{s}}\bar{x}_s^{-1} Cov(\bar{y}_s, \bar{y}_{\bar{s}}) \quad (3)$$

where T is the ratio unbiased predictor $\bar{y}_s(\bar{x}_P/\bar{x}_s)$. The optimal sample is that minimising (3) and might not necessarily include the n largest units (strategy F). Strategies B, C, D and F (based on a *purposive* sample selection) have been compared with A and E (evaluated on the basis of 1.000 random samples replications), where y is the monthly turnover for each of the first 10 months in 2006 and x is the average turnover 2004-2005. For method F, the sample selection was carried out choosing the sample leading to the lowest level of (3) among the same 1.000 replications used for methods A and E. Separate simulations have been carried out inside 8 strata: not specialised retail trade (employment classes 6-9, 10-19, >19) and specialised retail trade (1-2, 3-5, 6-9, 10-19, >19). Table 1 reports the mean of absolute percent errors (*MAPE*) for 2 sample coverages (25% and 50%), got averaging results concerning months and strata: results for the total derive from a simple or a weighted mean based on the true weights used in the survey. The sample mean is always the best estimator (bold) for specialised shops, but for those not specialised it can be always improved by strategy C (ratio estimator and purposive sampling). Due to the larger weight of the latters, strategy C could reduce total MAPE: with $f=25\%$, from 3,18% to 1,36% - as well as strategy E, that is quite always second best (underlined); with $f=50\%$, from 2,04% to 0,70%.

Table 1: Mean absolute percent errors (*MAPE*) for different domains and strategies

Domain	N	MAPE with $f=25\%$						MAPE with $f=50\%$					
		A	B	C	D	E	F	A	B	C	D	E	F
Simple mean of strata													
Specialised	162	1,93	6,12	4,66	4,03	<u>3,94</u>	6,33	0,63	2,32	<u>1,59</u>	1,72	2,29	3,25
Not specialised	162	2,76	2,14	1,25	1,72	<u>1,26</u>	2,63	2,04	1,13	0,46	1,39	<u>0,56</u>	1,89
Total retail trade of food	324	2,24	4,63	3,38	3,17	<u>2,94</u>	4,94	1,16	1,88	<u>1,17</u>	1,60	1,64	2,74
Weighted mean of strata													
Specialised	162	2,08	3,74	3,17	3,30	<u>3,07</u>	5,41	0,66	<u>1,39</u>	1,53	2,01	2,07	2,63
Not specialised	162	3,50	1,91	0,83	1,65	<u>0,87</u>	3,30	2,44	1,20	0,46	1,90	<u>0,65</u>	2,61
Total retail trade of food	324	3,18	2,32	1,36	2,02	<u>1,36</u>	3,77	2,04	1,25	0,70	1,92	<u>0,97</u>	2,62

References

- Cicchitelli G., Herzel A., Montanari G.E. (1992), *Il campionamento statistico*. Il Mulino, Bologna.
- Gismondi R. (2007), "Quick Estimation of Tourist Nights Spent in Italy", *Statistical Methods and Applications*, 2007-I, on-line publication, Springer & Verlag.