

Outlier definition problems in short term business surveys

*Problemi connessi alla definizione dei valori anomali
nelle indagini congiunturali sulle imprese*

Nicoletta Cibella

Angela Claudia Seeber

Istat, via Balbo, 16, Roma, seeber@istat.it

Riassunto: Con riferimento alle indagini congiunturali sul fatturato delle imprese, ci si pone l'obiettivo di studiare l'effetto che campionamento, mancata risposta e diversi criteri di ineleggibilità hanno sull'indice di riferimento, dove possono introdurre distorsione rispetto alla vera variazione tra fatturati totali. Per quanto riguarda i criteri di esclusione basati sul concetto di valore anomalo, sorge un ulteriore problema: il troncamento delle distribuzioni asimmetriche comporta il rischio di un indice di riferimento difficilmente interpretabile.

Keywords: outlying values, ineligible values, turnover index

1. Introduction

The aim of this paper is to provide some insight on how the exclusion of not observed or ineligible population units affects the value of the corresponding reference index. For this purpose, we calculated the annual turnover index at the population and different observation levels, using different eligibility criterions. Our data sources were the whole sale trade part of the Italian annual business registers 2003-04, combined with 2003-04 observation level indicator variables from the Istat whole sale trade short term business survey, whose objective is to estimate quarterly and annual year-to-year turnover indexes. Reflecting real sample and estimation procedures, we used the same functional index form for all levels, namely a panel index, defined as a weighted mean of stratum 2003-2004 year-to-year indices, which are calculated on enterprises satisfying particular eligibility criterions for 2003-2004. Strata are obtained by crossing 9 kinds of economic activity, three size classes in terms of employees and three geographic areas, referred to 2003. The fixed stratum weights are the registered population turnover quotes in 2003.

2. Methods and results

For each observation level and eligibility criterion we calculated the weighted target index, the weighted geometric index and statistics referring to the distribution of log-residuals to stratum means. All distributional analyses were performed on logarithmic annual variations, considering that original changes are highly skewed (Karlberg, 2000). As for observation levels we examined the registered population present since sample base year 2000 (P), the current theoretic survey sample (TS), sample units with at least one observed quarterly year-to-year ratio (OS), units with all four quarterly ratios

observed (COS), units with at least one preliminarily observed ratio (POS), units with all four ratios preliminarily observed (CPOS). As for population ineligibility criteria we considered registered deaths (including society fusions or splits), stratum changes, as well as tail cut rules, determined by analyses of stratum residuals. At the population level, excluding ceased units, only a few extreme outlying units were detected, which however did not affect sample statistics being none of them a survey respondent. After this restriction of the reference population, observation levels continued to be not ignorable with respect to index values and log-residuals, inducing decreased standard deviations as well as increased leptocurtosis and negative skewness. Only when units with stratum changes were deleted, yielding changed but similar geometric index values and almost symmetrical log-residuals at all observation levels, the strong observation level effects on target index values could be removed by applying a cut-off equal to five times the standard deviation of population log-residuals (table 1). More generally, the application of different tail cut rules yielded almost unchanged stratum change restricted sample indices, while corresponding population indices lowered significantly down, indicating that survey respondents represent the regular body of the reference population, where not observed values might be errors or true values.

Table 1: *Indices and log-ratio residuals for stratum change restricted subsets*

	OUTLIER RESTRICTED							TAIL CUT RESTRICTED						
	N	Index	Geo-index	Log-mean	Stratum log-residuals			N	Index	Geo-index	Log-mean	Stratum log-residuals		
					Std	Ske	Kur					Std	Ske	Kur
P	246.232	6,0	1,6	-0,02	0,47	-0,9	14,9	244.657	3,1	1,9	-0,01	0,40	-0,5	7,1
CT	5.237	8,0	0,6	-0,01	0,37	-1,6	17,1	5.214	3,2	1,0	-0,01	0,30	-1,4	13,4
OS	4.548	8,5	1,2	-0,00	0,29	-0,5	38,1	4.540	2,8	1,0	-0,00	0,26	-1,5	14,3
POS	4.406	10,0	1,4	-0,00	0,30	0,1	31,3	4.397	3,0	1,2	0,00	0,26	-1,5	14,2
COS	3.389	3,1	2,0	0,01	0,24	-0,4	44,6	3.386	3,1	2,0	0,01	0,22	-0,3	9,6
CPOS	2.463	2,9	2,0	0,02	0,22	0,0	26,8	2.462	2,9	1,9	0,02	0,20	-0,7	11,2

These results reflect interactions between the choice of illegibility criteria and observation level effects and, more generally, the fact that the implicit estimation model assumes that the expected year-to-year ratios depend on stratum variables only. Especially when true outlying sample values are deleted, the corresponding reference index might thus be better estimable, depending however not only on the observed sample size but also on violations of the ignorability hypothesis regarding observation levels with respect to the estimation model. Similarly, if units with structural changes are deleted, the corresponding reference index might be more accurately estimated, but of less interest to users. If deletion of sample units from estimates cannot be avoided, exclusion criteria should therefore be based on fixed definitions, which should be made accessible to users, rather than on current sample based statistical rules. On the other hand, better estimation models might be identified in order to reduce observation level biases with respect to both structural and turnover changes. Finally we should consider that in survey practice exclusion criteria and non response refer to quarterly data and estimates, which implies the need of annual pendants to explicit quarterly exclusion rules, at least when register reference data is used in order to investigate bias sources.

References

Karlberg F. (2000) Survey Estimation for Highly Skewed Populations in the presence of zeroes, *Journal of Official Statistics*, Vol.16, No.3, 229-241.