

A focus on a new approach to weighting in official sample surveys: the smoothed estimator

Veronica Rondinelli, Emanuela Scavalli ¹

Abstract:

The variability of estimators often depends on many elements. In this work a new approach of estimation, proposed by Beaumont, is presented. It is based on smoothed estimators, which have the aim to reduce the variability of unstable weights. An application on these estimators to the ISTAT Farm Business Survey (RICA-REA) 2006 is presented; this survey is affected by a high variability on estimators, especially at regional level. The obtained results are compared to those obtained applying the traditional estimators.

Keywords: smoothed weights, design-based estimators, sample design and estimation

1. Introduction

The validity of estimates and of their inference process depends on the model underling the definition of sample weights. The estimates of the parameters of interest are influenced by the sample design variables; this means that it exists correlation among those and the produced estimates. As a consequence, the estimates of population totals which are weakly related with the design variables could be inefficient. The Beaumont approach (Beaumont, 2008), named *smoothed approach*, has the purpose of increasing the efficiency of each design-based estimator, focusing the attention on a new estimator which takes into account of new sample weights obtained throughout a model which relates some variables of interest with the original weights.

While in the traditional model-based approaches it is necessary to postulate as many hypotheses and models as variables of interest are, the advantage of the smoothed approach is in applying only a single model to the weights.

The smoothed approach should reduce the variability of the estimators by reducing the variability of the weights throughout their modelisation. These new estimators are consistent and more efficient than the model-based estimators; moreover, they present lower variance even though they could introduce bias.

In this work we present an application of this approach to the ISTAT survey on the Farm Business in 2006. The obtained results are compared with those obtained with the weights calculated by traditional methods (Falorsi et al., 1998).

¹ Veronica Rondinelli
ISTAT, via C. Balbo 16 Rome, veronica.rondinelli@istat.it

Emanuela Scavalli
ISTAT, via C. Balbo 16 Rome, <mailto:scavalli@istat.it>

2. The smoothed estimator

We suppose to estimate the vector of population totals $T_y = \sum_{k \in U} y_k$ where y_k is the vector of variables of interest for the unit k -th and U is the finite population of size N .

We denote Y the matrix of N -rows containing y_k in the row k -th. We select a sample s of size n according to the probabilistic sample design $p(I|Z, Y) = p(I|Z)$ where Z is the N -rows matrix containing z_k in the k -th row, z_k being the vector of design variables for the population unit k -th and $I' = (i_1, \dots, i_N)$ being a vector of sample inclusion indicators where $i_k = 1$ if unit k -th is selected into the sample s , $i_k = 0$ otherwise.

We define generalised design-based inference any inference that is conditional to Y but not to I . In the smoothed approach we consider the inference related to the distribution $F_{I,Z|Y}$ which is a particular case of the generalised inference. With this type of inference only I is viewed as being random.

In the design-based theory the natural estimator of T_y is the unbiased Horvitz-Thompson estimator (HT)

$$\hat{T}_y^{HT} = \sum_{k \in s} w_k y_k$$

where $w_k = 1/\pi_k$ is the sample weight of unit k -th and $\pi_k = E(i_k | Z, Y) = E(i_k | Z)$ is its inclusion probability.

The stronger the variables of interest are related to the design variables (for example in stratification design), the more efficient the HT estimator is. However, surveys are in general multi-purpose and stratification is strongly related only to one or few variables of interest but less related to the other variables.

To solve this problem we consider the smoothed variable (SM)

$$\tilde{T}_y^{SHT} = \sum_{k \in s} \tilde{w}_k y_k,$$

where $\tilde{w}_k = E(w_k | I, Y)$ is a *smoothed weight* for unit k .

The idea is to reduce the weights variability by removing their noise considering their conditioned expected value. This variable $\tilde{T}_y^{SHT}$ is not an estimator since it depends on the unknown weights $\tilde{w}_k$. For this reason we modelize the design weights w_k in order to obtain the estimator $\hat{w}_k$ of $\tilde{w}_k = E(w_k | I, Y)$. At this point we can construct the HT *smoothed estimator*

$$\hat{T}_y^{SHT} = \sum_{k \in s} \hat{w}_k y_k.$$

To estimate the *smoothed weight*, we assume, for $k \in s$, that $\tilde{w}_k = g_s(y_k)$, underling the fact that function g_s varies from one sample to another. Many types of models can be defined (Pfeffermann and Sverchkov, 1999) and below we describe the two of them which have been selected among those more suitable to our data.

The first model assumes

$$w_k = h_k' \beta + v_k^{1/2} \varepsilon_k, \quad (1)$$

where ε_k given I and Y are errors independently and identically distributed with $E(\varepsilon_k | I, Y) = 0$ and $\text{var}(\varepsilon_k | I, Y) = \sigma^2$, β and σ^2 are unknown model parameters and the vectors h'_k and $\nu_k > 0$ are known functions of y_k .

In the first model the *smoothed weights* are defined as $\tilde{w}_k = h'_k \beta$ and they are estimated by $\hat{w}_k = h'_k \hat{\beta}$ where $\hat{\beta} = (\sum_{k \in S} \frac{h'_k h'_k}{\nu_k})^{-1} \sum_{k \in S} \frac{h'_k w_k}{\nu_k}$ is an estimator of β obtained using the generalised least-squares method. Two extreme cases of the first models are when:

- weights are independent from the design variables, in this case the model becomes $w_k = \beta + \varepsilon_k$, that is $h'_k = 1$ and $\nu_k = 1$ and we obtain $\hat{w}_k = \hat{N}/n$, $\hat{N} = \sum_{k \in S} w_k$. The HT smoothed estimator becomes $\hat{T}_y^{SHT} = \hat{N} \sum_{k \in S} y_k n$;
- the variables of interest predict perfectly weights, therefore $w_k = h'_k \beta_k$ and, in this case, the smoothed estimator coincides with the HT estimator with no gain in efficiency.

In practice we have an intermediate situation between these two extreme situations.

A second model assumes

$$w_k = 1 + \exp(h'_k \beta + \nu_k^{1/2} \varepsilon_k), \quad (2)$$

where the *smoothed weights* are defined as $\tilde{w}_k = 1 + \exp(h'_k \beta) E\{\exp(\nu_k^{1/2} \varepsilon_k) | I, Y\}$. This model ensures that weights assume values greater than 1.

These are only two examples of model which could be used, but, obviously, many more models exist and the choice among them depends on the type of data which we are examining.

The estimation of the variance of the *smoothed* estimator can be obtained or by developing the formula according to the related theory, or by using methods of Taylor linearization for not linear functions and re-sampling techniques (McCarthy, 1966, 1969). As far as the estimates of bias regards, Gwet and Rivest suggests the formula

$$\hat{B}^2 = \max\{0, (\hat{T}^{SM} - \hat{T}^{HT})^2 / \hat{\text{var}}(\hat{T}^{SM} - \hat{T}^{HT})\} \quad (\text{Gwet and Rivest, 1992}) \quad (3)$$

where $\hat{\text{var}}(\hat{T}^{SM} - \hat{T}^{HT})$ is the estimators of the variance of the difference of the two estimators.

Therefore, the mean square error could be estimated by

$$M\hat{S}E(\hat{T}^{SM}) = \min\{\hat{\text{var}}(\hat{T}^{SM}) + B^2, \hat{\text{var}}(\hat{T}^{HT})\} \quad (4)$$

where $\hat{\text{var}}(\hat{T}^{SM})$ and $\hat{\text{var}}(\hat{T}^{HT})$ are the estimators of the variances respectively of the smoothed and of the Horvitz-Thompson estimators.

3 An application to the Farm Business Survey

The smoothed approach has been applied to the ISTAT Farm Business Survey (RICA-REA) 2006. The survey furnishes a complete and exhaustive picture of the economic situation of the Italian agriculture through the collection of variables (REA and RICA) on the economic results of farms. The sample strategy for the 2006 survey uses a random sample design stratified by region, typology and business size (defined by the unit of economic dimension and by the working days). The business size has been defined in different ways for each region. The allocation of the sample units among strata is such to minimise the planned errors at national (0.03) and regional (among 0.05 and 0.10) level for some strategic variables: income gross standard, gross production, intermediate costs, amortizations, contributions, totals costs, employed labour costs and production (basic prices).

Since the stratification has been defined at regional level, we make the application on data related to each region.

The relationships among weights (with non-response correction and calibrated to some known totals) and some main variables (costs, production, value added ...) have been studied for the sample of farms of each region. Then, some models have been defined:

$$\text{weights} = g_1(\text{production}), \text{weights} = g_2(\text{costs}), \text{weights} = g_3(\text{value added}), \dots,$$

and they have been used in the construction of the smoothed estimator.

Therefore, the vectors of fitted values have been used for the construction of the *smoothed weights* defined in the first model (1) and a calibration on some known external benchmarks has been made (Deville and Särndal, 1992). Then, estimators of some variables of interest for the survey and their relative errors have been calculated. The variances of such estimators have been calculated using the Balanced Repeated Replications technique (Zannella, 1989) and the formula (3) and (4). It was also done an application by using the second model (2), but the obtained weights resulted extremely high showing the inadequacy of this model to this specific survey.

For some variables of interest (production, added value, costs, employed labour costs, working days, contributions, social contributions, revenues), some measures of variation and efficiency have been calculated, in particular:

- the coefficients of variation for the traditional HT estimator (CV_{HT}) and for the smoothed HT estimator (CV_{SM}),
- the relative distance of the estimates $RD = 100(\hat{Y}^{SM} - \hat{Y}^{HT})^2 / \hat{Y}^{HT}$,
- the statistic of WALD = $(\hat{Y}^{SM} - \hat{Y}^{HT})^2 / \text{var}(\hat{Y}^{SM} - \hat{Y}^{HT})$ that gives a measure of the design bias,
- the relative efficiency given by $RE = 100(\text{var}(\hat{Y}^{HT}) / \text{mse}(\hat{Y}^{SM}))$.

The results reported in tables 1 to 5 show that in general the *smoothed* estimator compared to the Horvitz-Thompson one produces estimates with less variability in almost all Italian regions, with the exception of Valle D'Aosta, Marche, Puglia and Veneto (Table1). It should be noted that, for example, in Veneto coefficients of variation of the Horvitz Thompson estimator are lower than those of the smoothed estimator for whole variables except for the employed labour costs in which the Horvitz-Thompson estimator is 4.4 percent points more than smoothed one. On the

other hand, in three regions (Abruzzo, Lombardia and Toscana) the smoothed estimator presents on whole variables considered lower coefficients of variation, while in the remaining regions some variables have a gain in efficiency with the smoothed estimator and other regions do not present significant differences among the two estimators.

As far as variables of interest are concerned, the employed labour costs and the revenues have in all regions a great gain in efficiency by using the smoothing approach. From all that we can deduce the strong variability which characterizes the estimators of economic variables in the primary sector at the regional level, variability that is inside the agricultural data.

The results are also confirmed in Table 3 where the relative efficiency (RE) is always greater than 100, pointing out that the smoothed method is more efficient than the traditional one. This emphasizes a strong reduction of the estimates variability in the smoothed approach.

However, Table 4 underlies that the smoothed method introduces always bias, in some cases negligible, but in other cases important. The highest value (55.06) is observed in Liguria on the variable social contributions. However, there are unbiased estimators, for example the costs in Sardegna or the contributions in Trento.

By observing the relative difference (RD) of the estimates we can note, in general, lower values for the smoothed estimates than those obtained with the traditional approach (Table 5.). However, an exception is represented by the Valle d'Aosta and Marche in which the difference is positive.

Table 1. *Coefficients of variation for traditional (HT) and smoothed (SM) estimators*

	production		added value		costs		employed labour costs	
	HT	SM	HT	SM	HT	SM	HT	SM
Piemonte	16.62	5.17	16.05	5.38	17.91	8.89	41.60	21.67
Valle d'Aosta	12.98	16.78	14.89	19.00	11.94	15.38	43.11	48.77
Lombardia	8.37	6.37	7.76	5.69	11.04	9.61	14.15	7.31
Veneto	8.81	8.98	10.46	11.34	9.94	10.40	10.82	6.40
Friuli Venezia Giulia	9.29	7.93	6.83	9.30	13.10	8.79	18.65	14.15
Liguria	3.85	3.78	3.91	3.87	4.32	4.18	12.43	8.37
Emilia Romagna	6.31	5.78	8.32	10.66	5.98	5.55	15.79	10.10
Toscana	6.67	5.49	7.11	5.71	7.08	6.02	11.10	6.32
Umbria	7.12	6.96	8.63	8.71	7.07	6.15	15.12	7.63
Marche	5.48	6.50	7.34	8.67	5.93	6.74	26.02	18.53
Lazio	5.54	5.31	6.00	5.34	5.62	6.42	19.52	6.71
Abruzzo	5.53	3.92	6.68	4.58	4.64	3.82	18.28	11.59
Molise	6.89	7.19	7.88	8.11	6.65	7.67	16.15	12.94
Campania	11.51	8.34	13.09	16.11	8.75	11.67	11.95	5.41
Puglia	4.49	4.63	5.48	5.59	4.27	3.92	6.77	6.79
Basilicata	6.77	6.09	7.46	7.39	6.79	4.92	14.78	8.50
Calabria	18.99	16.46	19.96	17.53	14.48	12.27	21.57	17.75
Sicilia	5.40	4.91	5.69	4.83	5.73	5.86	8.04	3.81
Sardegna	7.16	4.04	6.42	4.32	9.52	5.14	26.70	7.11
Trento	8.93	8.66	7.74	6.76	11.84	11.63	19.09	14.72
Bolzano	4.44	3.81	4.31	4.08	5.74	4.24	9.35	9.15

Table 2. *Coefficients of variation for traditional (HT) and smoothed (SM) estimators*

	working days		contributions		social contributions		revenues	
	HT	SM	HT	SM	HT	SM	HT	SM
Piemonte	4.71	1.82	4.32	7.62	11.15	2.82	17.61	5.45
Valle d'Aosta	9.35	9.35	6.22	8.90	19.30	23.56	10.44	14.61
Lombardia	4.71	4.42	12.71	7.98	5.75	4.47	8.50	6.81
Veneto	2.83	3.05	7.42	9.73	4.44	2.99	8.76	8.83
Friuli Venezia Giulia	4.49	2.50	13.74	9.33	8.66	4.35	9.27	8.20
Liguria	2.41	2.70	15.84	15.15	3.70	3.82	4.24	4.03
Emilia Romagna	3.89	1.32	7.93	16.45	8.42	4.06	6.13	5.69
Toscana	5.38	4.60	10.72	9.98	8.73	8.08	6.43	5.35
Umbria	5.62	5.65	8.76	7.29	6.83	7.15	7.16	6.41
Marche	6.46	7.49	7.77	8.00	8.81	8.40	5.69	6.85
Lazio	3.37	3.39	9.89	12.69	6.85	5.47	5.77	5.58
Abruzzo	4.70	4.08	5.53	5.13	5.37	5.38	5.97	4.17
Molise	5.45	4.82	5.57	5.76	6.39	6.44	7.22	7.55
Campania	4.52	5.72	7.96	8.59	7.03	4.68	12.06	6.11
Puglia	4.04	4.81	5.24	4.78	4.44	4.85	4.68	4.65
Basilicata	4.75	4.77	7.95	7.54	13.16	12.61	7.04	6.02
Calabria	5.35	4.87	27.90	26.81	18.23	15.22	19.36	16.75
Sicilia	3.36	3.55	6.63	5.62	5.43	5.53	5.48	5.09
Sardegna	8.76	9.55	5.85	5.68	13.47	4.06	7.78	4.30
Trento	5.47	4.15	18.75	32.50	5.66	5.22	9.96	9.69
Bolzano	4.01	3.40	15.91	14.98	5.34	4.95	4.69	4.31

Table 3. *Relative efficiency (RE) of the estimates*

	production	added value	costs	employed labour costs	working days	contributions	social contributions	revenues
Piemonte	161	1.198	100	114	104	100	102	230
Valle d'Aosta	100	100	100	100	114	102	100	100
Lombardia	210	207	159	100	101	129	100	194
Veneto	119	115	106	100	100	121	100	122
Friuli Venezia Giulia	101	112	100	100	100	306	100	102
Liguria	100	100	100	233	100	118	100	100
Emilia Romagna	100	100	135	100	100	100	100	104
Toscana	100	100	100	100	100	119	100	100
Umbria	133	121	107	100	144	100	108	170
Marche	102	102	100	100	100	105	100	102
Lazio	129	144	117	100	100	124	100	125
Abruzzo	206	216	161	224	179	113	100	215
Molise	110	108	102	156	124	112	102	106
Campania	101	100	108	100	138	100	100	201
Puglia	136	131	158	100	105	128	116	144
Basilicata	138	127	103	103	100	115	100	150
Calabria	135	136	130	130	123	118	131	135
Sicilia	100	100	102	100	100	173	100	100
Sardegna	342	221	378	643	106	100	157	364
Trento	126	108	224	344	217	100	176	168
Bolzano	100	100	100	128	118	158	100	100

Table 4. *WALD statistic*

	production	added value	costs	employed labour costs	working days	contributions	social contributions	revenues
Piemonte	1.78	0.84	2.80	2.49	6.09	4.69	3.62	1.57
Valle d'Aosta	2.54	1.90	2.49	1.46	0.28	0.01	0.73	4.08
Lombardia	0.92	0.99	0.70	4.42	2.44	2.03	3.37	0.76
Veneto	0.36	0.52	0.10	9.25	20.81	0.07	18.93	0.46
Friuli Venezia Giulia	5.39	0.36	9.40	9.67	20.97	3.08	19.36	5.25
Liguria	8.24	5.53	18.38	2.59	42.88	4.70	55.06	8.06
Emilia Romagna	3.07	3.01	0.77	11.68	20.06	0.11	13.72	1.80
Toscana	13.61	11.67	12.28	27.71	15.29	0.10	24.64	15.70
Umbria	0.47	0.07	2.43	7.80	0.55	7.29	0.08	0.74
Marche	0.92	0.06	0.88	9.00	0.01	0.15	13.21	0.81
Lazio	0.41	0.92	0.01	3.73	17.59	0.03	16.91	0.15
Abruzzo	0.00	0.04	0.21	0.87	1.31	1.59	9.30	0.10
Molise	0.66	0.00	1.83	2.15	2.03	0.06	2.56	1.35
Campania	4.39	6.62	0.00	14.07	1.68	18.61	10.05	3.10
Puglia	1.07	1.51	0.48	16.99	0.17	0.30	0.39	1.25
Basilicata	1.04	0.23	4.10	7.80	8.58	0.00	9.89	1.11
Calabria	0.03	0.24	0.72	1.19	0.04	4.11	1.11	0.01
Sicilia	4.86	6.64	1.74	52.75	19.23	0.01	27.99	4.79
Sardegna	0.03	0.33	0.00	1.30	1.58	5.97	2.00	0.02
Trento	3.23	6.06	0.04	5.11	2.09	0.00	0.48	2.90
Bolzano	11.05	10.42	12.61	0.47	2.43	0.39	18.66	12.42

Table 5. *Relative distance (RD) of the estimates*

	production	added value	costs	employed labour costs	Working days	contributions	social contributions	revenues
Piemonte	-20.49	-13.15	-27.57	-56.91	-8.88	-13.89	-18.16	-20.28
Valle d'Aosta	10.23	10.14	10.96	18.62	2.53	0.76	6.70	13.55
Lombardia	-5.03	-5.38	-4.49	-22.58	-4.10	-12.37	-7.11	-4.40
Veneto	-4.55	-7.69	-1.84	-24.51	-8.35	-1.76	-14.76	-5.06
Friuli Venezia Giulia	-15.82	-4.11	-25.57	-47.93	-19.54	-15.06	-31.97	-15.37
Liguria	-5.08	-4.58	-6.39	-10.52	-4.37	-14.00	-8.86	-6.02
Emilia Romagna	-10.39	-14.84	-5.70	-44.68	-15.86	5.65	-20.26	-8.13
Toscana	-14.75	-14.69	-14.82	-40.04	-8.56	1.91	-16.84	-15.84
Umbria	-3.94	-1.87	-7.24	-37.07	-2.78	-15.47	-1.05	-4.97
Marche	2.00	0.68	2.68	29.64	0.38	0.90	-4.90	1.86
Lazio	-2.47	-4.03	0.39	-34.59	-11.14	1.26	-22.33	-1.60
Abruzzo	-0.15	0.72	-1.44	8.03	-2.78	3.65	-4.95	-1.06
Molise	2.19	-0.15	6.03	-9.34	-3.07	0.69	-3.75	3.56
Campania	-23.15	-31.23	0.14	-34.87	-5.52	28.16	-18.13	-17.59
Puglia	-3.96	-4.77	-2.93	-17.44	-1.85	2.44	-2.46	-4.15
Basilicata	-3.10	-1.68	-6.33	-21.28	-5.09	0.15	-10.00	-3.44
Calabria	-0.80	-2.23	3.59	5.73	0.44	-9.95	3.98	-0.54
Sicilia	-12.54	-14.22	-8.45	-47.69	-11.81	0.42	-20.15	-12.86
Sardegna	1.11	2.67	-0.29	-27.59	-5.84	8.08	-17.01	0.98
Trento	-8.93	-13.38	-0.96	-21.47	-6.40	-0.53	-3.24	-9.01
Bolzano	-10.95	-9.44	-15.50	-3.05	-5.48	-7.69	-14.79	-11.58

4 Conclusions

In conclusion, although there is an extreme variability of the phenomena in the farms it is particularly complex to define a stratified design strictly correlated with all the variables of interest. The smoothed approach presents different results. In some regions smoothed approach is better because the coefficients of variation are low. In other regions the traditional approach seems to be preferred because the smoothed estimator introduces high bias in the estimates. However, in general the smoothed approach seems to presents good results, or at least results quite similar to the traditional approach. Only in few cases it produces a significant worsening of estimates.

The difficulty of apply such approach mainly consists of seeking and defining the models that relates the weights with some explicative variables: the problem is mainly in the fact that they are related to some variables highly correlated only with some variables of interest, for which the advantage in terms of estimates is positive, but for the other variables not correlated it introduces a worsening of the estimates. In any case, the results of the application encourage the use of this new approach.

References

- Beaumont J.F. (2008) A new approach to weighting and inference in sample surveys, *Biometrika*, 95, 3, pp.539-553.
- Deville, J.C. e Särndal, C. E., (1992) Calibration Estimation in Survey Sampling, *Journal of the American Statistical Association*, Vol. 87, No. 418, 376-382
- Guwet J.P., Rivest L.P (1992), Outlier resistant alternatives to the ratio estimator, *Journal of the American Statistical Association*, 87, 1174-1182
- McCarthy P.J.(1966) Replication: an approach to the analysis of data from Complex surveys, Vital and Health statistics, Public Health Service, Washington
- McCarthy P.J.(1969) Pseudoreplication: Half-samples. *Review of the International Statistical Institute* 37, 239-264
- Falorsi P.D., Ballin M., De Vitiis C., Scepi G., (1998) Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'Istat, *Statistica Applicata*, Vol.10, No. 2,
- Pfeffermann D., Sverchkov M. (1999), Parametric and semi-parametric estimation of regression models fitted to survey data, *Sankhya*, Series B, 61, 166-186
- Zannella F. (1989), Manuale di tecniche di indagine: tecniche di stima della varianza campionaria, *Note e Relazioni*, n.1 ISTAT