

Some Results on Probabilistic Expert Systems for Finite Population Sampling

*Alcuni risultati sui sistemi esperti probabilistici per
il campionamento da popolazioni finite*

Marco Ballin¹, Mauro Scanu¹, Paola Vicard²

¹ Dipartimento della Produzione Statistica, ISTAT, Roma
e-mail: scanu@istat.it

² Dipartimento di Economia, Università Roma Tre

Riassunto: La scelta della strategia campionaria disegno/stimatore privilegia l'uso di variabili ausiliare connesse con le variabili di interesse. Le variabili ausiliare sono utilizzate per definire un nuovo sistema di pesi campionari e/o uno stimatore opportuno. Ballin *et al.* (2005) mostrano come gli usuali stimatori implicitamente fanno riferimento a una struttura di dipendenza particolare fra variabili di interesse e ausiliarie: il modello saturato. In questo lavoro si discutono alcune proprietà degli stimatori suggeriti da strutture di dipendenza fra variabili rappresentate mediante sistemi esperti probabilistici.

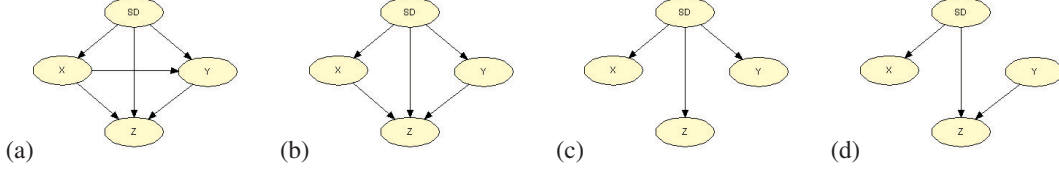
Keywords: complex sampling design, graphical models, Bayesian networks.

1. Introduction

Efficiency of a finite population sampling strategy, i.e. of the pair estimator/sampling design, is strictly linked to the concept of probabilistic relationship between variables. For instance, a stratified design is appropriate when stratification and target variables are dependent, as well as the ratio or regression estimators are desirable when auxiliary variables are connected with the target ones. In a multivariate context the dependence relationship between the target variables and the possible auxiliary variables can be rather complex: marginal or conditional independences between these variables should be taken into account. For the sake of simplicity, our analysis is restricted to categorical variables, and the aim is the estimation of a relative frequency contingency table. Ballin *et al.* (2005) define a class of estimators based on the dependence structure of all the involved variables and focus on the dependence relationships represented with a class of graphical models known as probabilistic expert systems (PES; also called Bayesian networks, see Cowell *et al.*, 1999). This class of models is justified by the following aspects: (i) if a dependence model is a PES, the estimator is easy to compute (Section 2); (ii) PES properties and features, as the information propagation algorithm, can be exploited for describing poststratification procedures (cfr. Ballin *et al.*, 2005); (iii) tools for imputing missing items developed in a PES context (cfr. Di Zio *et al.*, 2005) can be used also in the context of complex survey sampling.

This paper aims at illustrating estimators of contingency tables assisted by knowledge on the dependence structure of the variables under study (Section 2), and some aspects of their efficiency (Section 3).

Figure 1: 4 possible PES structures for X, Y, Z and the design variable SD .



2. Estimators assisted by PES structure

Let X, Y , and Z be three categorical variables with joint frequency distribution (f.d.) on a finite population of N units $\{\theta_{xyz} = N_{xyz}/N\}$, which is the target. Let $\mathcal{S}$ be a sample drawn from the population according to a complex survey design, and $\omega_i, i \in \mathcal{S}$, be the sample weights assigned to the sampled units by the survey design. Ballin *et al.* (2005) define an additional variable, the sample design variable (SD), as that categorical variable whose categories are the different sample weights ($\omega_{(sd)}$), and f.d. is given by the fraction of the overall sample weight owned by the n_{sd} sample units with the same weights $\omega_{(sd)}$: $\theta_{sd} = \sum_{i \in s_{sd}} \omega_i / \sum_{i \in \mathcal{S}} \omega_i = n_{sd} \omega_{(sd)} / (\sum_{sd} n_{sd} \omega_{(sd)})$, where s_{sd} is the set of labels in $\mathcal{S}$ of those units with the same weight $\omega_{(sd)}$. For instance, in a stratified sample design, SD has as many categories as the number of strata and its f.d. is defined in terms of the overall strata weight. The strata are a combination of categories of archive variables, as the region of residence or the gender. In a complex setting, e.g. a multivariate contingency table estimation, it is not easy to define a sample design whose auxiliary variables are connected with all the target variables. It is important to assess what sampling strategy (design/estimator) should be chosen for any dependence relationship of (SD, X, Y, Z) . In this paper we held fixed the design and study the form of the estimator.

Ballin *et al.* (2005) focus on PES to represent the dependence model between the variables. In this way the form of the estimator can be derived using the *chain rule* (Cowell *et al.*, 1999). The chain rule gives a special factorization of the joint f.d. $\theta_{sd,x,y,z}^{(PES)} = \theta_{sd|pa(sd)} \theta_{x|pa(x)} \theta_{y|pa(y)} \theta_{z|pa(z)}$, where θ s are relative frequencies and $pa(\cdot)$ are the parents of the corresponding variables in the graphical structure (e.g. the chain rule for network (b) in Figure 1 is $\theta_{sd,x,y,z}^{(b)} = \theta_{sd} \theta_{x|sd} \theta_{y|sd} \theta_{z|sd,x,y}$). The following is a model assisted approach, i.e. PES should not hold strictly ($\theta_{sd,x,y,z} = \theta_{sd,x,y,z}^{(PES)}$ does not necessarily hold for any sd, x, y, z ; the previous equalities always hold when the PES is the complete network, see Figure 1 (a)). It suffices that the PES “fits” the population data, i.e. the PES is a simpler but adequate model well explaining the population data ($\theta_{sd,x,y,z} \sim \theta_{sd,x,y,z}^{(PES)} \forall sd, x, y, z$). The PES assisted estimator $\hat{\theta}_{xyz}^{(PES)}$ plugs the usual estimator defined by the sample design as a weighted average of the sampled units with weights given by the survey design, in each factor of $\theta_{sd,x,y,z}^{(PES)}$. When SD has no parents, the estimator has a particular form: $\hat{\theta}_{xyz}^{(PES)} = \sum_{sd} \theta_{sd} \hat{\theta}_{x|pa(x)} \hat{\theta}_{y|pa(y)} \hat{\theta}_{z|pa(z)}$ where θ_{sd} is known by design. In this case the factors $\hat{\theta}_{x|pa(x)}$, $\hat{\theta}_{y|pa(y)}$, and $\hat{\theta}_{z|pa(z)}$ in $\hat{\theta}_{xyz}^{(PES)}$ do not depend on the survey weights, either when the set of parents includes SD (e.g. $\hat{\theta}_{x|sd} = \sum_{i \in s_{sd}} I_x(x_i) / n_{sd}$), or not (e.g. $\hat{\theta}_x = \sum_{i \in \mathcal{S}} I_x(x_i) / n$). Furthermore $\hat{\theta}_{xyz}^{(PES)}$ has the usual form of a weighted mean, with weights $\omega_i, i \in \mathcal{S}$, if and only if the PES is the complete network, $\hat{\theta}_{xyz}^{(a)}$ for network (a) in Figure 1. Hence, this class of estimators enlarges the usual class of estimators by means of the PES structure.

3. Efficiency of the proposed estimators for large samples

Given that the PES just fits the population data, it can be easily proved that $\hat{\theta}_{xyz}^{(PES)}$ is neither unbiased nor consistent (letting the sample size to the population size) for θ_{xyz} . Only the usual one, $\hat{\theta}_{xyz}^{(a)}$, is always a consistent estimator of θ_{xyz} (and unbiased when $\sum_{i \in \mathcal{S}} \omega_i = N$, i.e. for the Horvitz-Thompson estimator). Nevertheless, experimental results (cfr. Ballin *et al.*, 2005) suggest that $\hat{\theta}_{xyz}^{(PES)}$ is more efficient than $\hat{\theta}_{xyz}^{(a)}$. Table 1 shows the results of a Monte Carlo evaluation of $\hat{\theta}_{xyz}^{(PES)}$: 4 finite populations of 10000 units have been generated from the 4 PES of Figure 1; from each population, $M = 500$ samples of size 1000 have been drawn according to a stratified sample design (SD has three categories); the corresponding 4 PES assisted estimators $\hat{\theta}_{xyz}^{(PES)}$ have been applied and their estimates have been compared to the finite population contingency table with: $RMSE(\hat{\theta}_{xyz}^{(PES)}) = \sum_{m=1}^M \sum_{xyz} [\theta_{xyz} - \hat{\theta}_{xyz}^{(PES)m}]^2 / (M\theta_{xyz})$, i.e. the mean of the relative mean square errors, where $\hat{\theta}_{xyz}^{(PES)m}$ is the PES assisted estimate of θ_{xyz} for sample $m = 1, \dots, M$. Let $\mathcal{G}$ be the correct PES for the population, and assume that $\mathcal{F}$ is the one used for the estimator. The diagonal of Table 1 (i.e. when $\mathcal{F} = \mathcal{G}$) shows the best Monte Carlo RMSE; the up right part of the table (i.e. for $\mathcal{F}$ with less dependencies - arrows - than $\mathcal{G}$) shows the largest RMSE; $\hat{\theta}^{(a)}$ is generally worse than $\hat{\theta}^{(PES)}$, unless the true model is (a). Note that being $\hat{\theta}^{(a)}$ unbiased, the gain in terms of RMSE of $\hat{\theta}_{xyz}^{(PES)}$ should be justified by a decrease in variance. A first attempt to justify the different RMSEs when there is a model misspecification can be done analyzing each factor defined by the chain rule in $\hat{\theta}_{xyz}^{(PES)}$. The chain rule for $\mathcal{G}$ and $\mathcal{F}$ identifies the set of variables with the same set of parents (say $V \in \mathcal{V}$) and its complement ($U \in \mathcal{U}$), where $\mathcal{U} \cap \mathcal{V} = \emptyset$ and $\mathcal{U} \cup \mathcal{V} = \{SD, X, Y, Z\}$. The factors for $V \in \mathcal{V}$ are identical for both $\hat{\theta}^{\mathcal{G}}$ and $\hat{\theta}^{\mathcal{F}}$. Each estimator $\hat{\theta}_{u|pa(u)}^{\mathcal{G}}$ is consistent (and under certain conditions unbiased) for $\theta_{u|pa(u)}^{\mathcal{G}}$. When $pa(u)$ in $\mathcal{F}$ is a subset of $pa(u)$ in $\mathcal{G}$, an independence or conditional independence that does not hold in the population is considered, and $\hat{\theta}_{u|pa(u)}^{\mathcal{F}}$ converges to $\theta_{u|pa(u)}^{\mathcal{F}} \neq \theta_{u|pa(u)}^{\mathcal{G}}$. On the contrary, when $pa(u)$ in $\mathcal{G}$ is a subset of $pa(u)$ in $\mathcal{F}$, $\hat{\theta}_{u|pa(u)}^{\mathcal{F}}$ is almost consistent, $\theta_{u|pa(u)}^{\mathcal{F}} \sim \theta_{u|pa(u)}^{\mathcal{G}}$, because the additional conditional variables are almost independent with U in the population given the other conditional variables. However, $\hat{\theta}_{u|pa(u)}^{\mathcal{F}}$ is more variable, given that $\theta_{u|pa(u)}^{\mathcal{F}}$ corresponds to a more complex cross tabulation than $\theta_{u|pa(u)}^{\mathcal{G}}$. For this reason, although unbiased and consistent, $\hat{\theta}_{xyz}^{(a)}$ has a larger RMSE when (a) does not hold (the corresponding factors of the chain rule are unnecessarily more complex). Note finally that, when $\hat{\theta}_{xyz}^{(PES)}$ is based on a PES with both an additional and a missing arrow, the lack of a parent is always predominant (e.g. see $\hat{\theta}^{(c)}$ for population (d) in Table 1). The previous considerations hold for each single factor of $\hat{\theta}_{xyz}^{\mathcal{F}}$. In fact, they cannot be automatically extended to the performance of the overall estimator $\hat{\theta}_{xyz}^{\mathcal{F}}$, for general $\mathcal{F}$ although Table 1 seems to confirm these results: this aspect deserves further research.

The previous results for the overall $\hat{\theta}_{xyz}^{\mathcal{F}}$ efficiency can be studied for each cell (x, y, z) . Figure 2 shows the ratio $RMSE(\hat{\theta}_{xyz}^{(a)})/RMSE(\hat{\theta}_{xyz}^{\mathcal{F}})$ for each (x, y, z) when (b), (c) and (d) are the correct PES. Also in this case, $\hat{\theta}_{xyz}^{\mathcal{F}}$ has a better performance than the usual $\hat{\theta}_{xyz}^{(a)}$ (all the ratios are larger than 1). Furthermore, this gain is more evident for the rarest cells. Again, this may be a result of factorizing the joint d.f. with the chain rule and simplifying

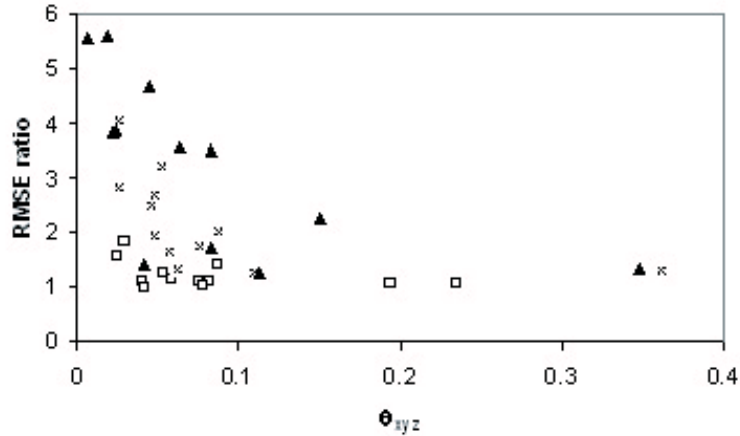
Table 1: RMSE of $\hat{\theta}_{xyz}^{\mathcal{F}}$, for 4 different populations with $\mathcal{G}$ given in Figure 1

Population ($\mathcal{G}$)	RMSE($\hat{\theta}_{xyz}^{(a)}$)	RMSE($\hat{\theta}_{xyz}^{(b)}$)	RMSE($\hat{\theta}_{xyz}^{(c)}$)	RMSE($\hat{\theta}_{xyz}^{(d)}$)
a	0.32	0.36	0.54	3.07
b	0.29	0.25	0.45	10.17
c	0.27	0.22	0.15	3.16
d	0.28	0.45	6.25	0.10

the dependencies in each factor by exploiting independence and conditional independence relationships. Finally the gain seems to be more evident for lighter structures (i.e. (c) and (d)) because each factor in $\hat{\theta}_{xyz}^{\mathcal{F}}$ is almost consistent but much less dispersed than the corresponding factor in $\hat{\theta}_{xyz}^{(a)}$.

In conclusion, PES assisted estimators are a class of estimators larger than the usual one in which a “best” estimator can be found. Note that, instead of the usual practice of fixing the form of the estimator as a weighted average, and finding the best set of weights, we fix the weights and find the best form of the estimator. Further studies will necessarily involve the choice of the overall sampling strategy, that should optimize both the aspects (weights and form of estimators).

Figure 2: RMSE ratios for $\hat{\theta}_{xyz}^{\mathcal{F}}$ and $\hat{\theta}_{xyz}^{(a)}$ when $\mathcal{F} = \mathcal{G}$ is (b), (c) and (d), $\forall (x, y, z)$.



References

- Ballin M., Scanu M., Vicard P. (2005) Model assisted approaches to complex survey sampling from finite populations using Bayesian networks, W.P. 54, Dipartimento di Economia (Università Roma Tre), <http://tinyurl.com/ajbuf>.
- Cowell R. G., Dawid A. P., Lauritzen S. L., Spiegelhalter D. J. (1999) *Probabilistic Networks and Expert Systems*, Springer, New York.
- Di Zio M., Sacco G., Scanu M., Vicard P. (2005) Multivariate techniques for imputation based on Bayesian networks *Neural Network World*, 2005/4, 303–309.