Bayesian Networks and Complex Survey Sampling from Finite Populations

Marco Ballin*, Mauro Scanu, Paola Vicard*****

National Institute of Statistics - Istat (* and **) and Università Roma Tre (***)
*via Adolfo Ravà 150, 00142 Roma, ITALY, ballin@istat.it
**via Cesare Balbo 16, 00184 Roma, ITALY, scanu@istat.it
***via Ostiense 139, 00154 Roma, ITALY, vicard@uniroma3.it

1 Introduction

We propose a novel methodology based on the concept of Bayesian network (BN, see Cowell *et al.*, 1999) for the estimation of a joint probability distribution of a set of categorical variables when samples are drawn according to complex survey designs. Note that, restricting ourselves to categorical variables, the previous aim corresponds to estimation of a contingency table, a very frequent problem in Offc ial Statistics.

BNs are graphical devices largely used in many different scientifc contexts, such as artifcial intelligence and multivariate statistics (Neapolitan, 2004). However, when estimating and using BNs, observations have always been considered as i.i.d. generations from a suitable joint distribution function. Up to now, BNs have never been defned and applied when sampling from fnite populations.

This paper shows that BNs can be easily adapted to the context of fnite survey sampling via the defnition of a suitable additional variable, in the following denoted with SD, representing the survey design. Hence, SD will be a categorical variable with as many states as the different inclusion probabilities of frst order. The BN representation allows the def nition of a much larger class of estimators, of the model assisted type (see Särndal *et al.*, 1992). Also, the possibility to use poststratifcation methods and, in general, integration of different surveys is illustrated.

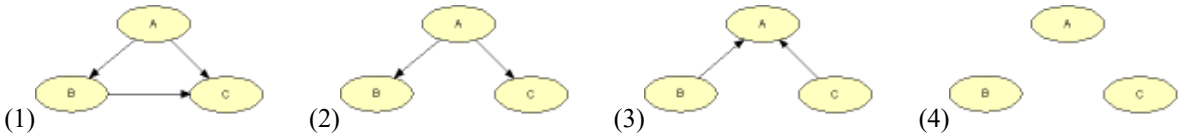
1.1 Bayesian networks

Despite the name, the term 'Bayesian' does not refer to the Bayesian inferential paradigm. A Bayesian network is just a graphical and numerical representation of a joint distribution of a set of variables, $(X_1, \ldots, X_k)$ say. Hence, a BN is the objective of the inference, which can be determined under either a likelihood based or a Bayesian procedure, see Neapolitan (2004) and references therein. The term Bayesian is due to an effci ent information propagation algorithm based on the Bayes theorem. This characteristic will be crucial, for instance, when applying poststratifcation (Section 2.4). A BN is characterized by: (i) a directed acyclic graph (DAG) showing the set of dependencies among variables and (ii) an inferential engine to make inference on the parameters of the model. A DAG is composed of nodes, each node representing a variable, and edges, each edge is an arrow linking a pair of nodes (for basics and defnitions on DAGs and BNs see for instance Jensen, 1996). Cycles are forbidden, in the sense that, following the direction of the arrows it is impossible to start from a node and end up in it. When two nodes X_i and X_j are connected by an arrow (i, j) pointing from X_i to X_j, the two nodes are probabilistically dependent and X_i is said to be a parent of X_j. Each node has attached the conditional distribution of the corresponding variable, say X_j, given its parents $pa(X_j)$. This representation allows the joint distribution of $(X_1, \ldots, X_k)$ to be factorized according to the dependencies shown in the DAG:

$$P(X_1, \ldots, X_k) = \prod_{j=1}^{k} P(X_j | pa(X_j)), \tag{1}$$

where $pa(X_j)$ can possibly be the empty set (in this case $P(X_j|pa(X_j)) = P(X_j)$). Once the BN has been estimated, its modular structure can be exploited to apply fast and effcient algorithms. For instance, the effect of changes in the distribution of some of the variables on the other variables can be easily computed (see Section 2.3). The interpretation of a BN in terms of the probabilistic relations among the variables can be described by the networks in Figure 1. Network (1), known also as 'complete network', implies that each variable is connected with the others. Network (2) shows independence of B and C given A. Network (3) shows marginal independence between B and C but conditional dependence between B and C given A. Network (4) shows independence between all the variables.

As already anticipated, both the structure of the BN (i.e. the set of edges) and the parameters of the distributions in (1) can be estimated under either a likelihood based or a Bayesian perspective. In the following, we will assume the structure of the BN as given. Given the structure, parameters will be estimated according to appropriate fnite population estimators. Note that we will



(1) (2) (3) (4)

Figure 1: Four possible network structures for the nodes (A, B, C). Note that redirections of some arrows in networks (1) and (2) produce equivalent joint distribution functions

not consider explicitly any model based assumption, hence the estimators cannot be justifed under either a likelihood based or a Bayesian approach. However, the parameter estimates will resemble those determined under maximum likelihood. This allows the use of usual software tools for computing probability distributions in BNs (as Hugin, `http:\\www.hugin.com`).

2. BN and fnite populations

Let us consider a f nite population of N units, and let $(x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, N$, be the values of k variables of interest, $(X_1, \ldots, X_k)$, of the population units. As already stated, we will restrict only to categorical variables. In this case, the joint distribution function corresponds to the relative frequency distribution computed on the population:

$$P(x_1, \ldots, x_k) = \frac{1}{N} \sum_{i=1}^{N} I_{x_1,\ldots,x_k}(x_{i1}, \ldots, x_{ik}), \tag{2}$$

where $I_{x_1,\ldots,x_k}(x_{i1}, \ldots, x_{ik})$ is the indicator function.
Assume that a random sample $\mathcal{S}$ of n units is drawn fro the population according to a sampling design that assigns a probability of inclusion π_i to each unit $i = 1, \ldots, N$. Let ω_i, $i \in \mathcal{S}$, be the fnal weight based on the sampling strategy (i.e. sampling design and estimator). The usual estimator of the joint distribution function (2) is:

$$\hat{P}(x_1, \ldots, x_k) = \sum_{i \in \mathcal{S}} I_{x_1,\ldots,x_k}(x_{i1}, \ldots, x_{ik}) \frac{\omega_i}{\sum_{i \in \mathcal{S}} \omega_i}, \tag{3}$$

henceforth the Direct Joint (DJ) estimator. Note that the DJ estimator is a ratio estimator, which corresponds to a Horwitz–Thompson estimator when $\sum_{i \in \mathcal{S}} \omega_i = N$. The DJ estimator can equivalently be rewritten with the help of a particular BN. This is just one of the possible estimators that the BNs can defne. In order to show all of them, it is necessary to highlight the role played by the weights ω_i in (2). Let SD be an additional categorical variable assuming as many states as the different survey weights, say H. Let $\omega_{(h)}$, $h = 1, \ldots, H$, be the SD states, s_h be the set of labels of the sample units with $\omega_i = \omega_{(h)}$, and n_h be the number of units in s_h. SD is associated to the marginal probability distribution given by the fraction of the total sample weight associated to the units in s_h:

$$P(SD = h) = \frac{\sum_{i \in s_h} \omega_i}{\sum_{i \in \mathcal{S}} \omega_i} = \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}}, \qquad h = 1, \ldots, H.$$

Given that information on the survey design is completely contained in SD, estimators computed given SD do not depend on the survey weights any more. For instance, the estimators of the marginal and conditional frequencies of a variable given SD are:

$$\hat{P}(X_u = x_u | SD = h) = \frac{\sum_{i \in s_h} I_{x_u}(x_{iu})}{n_h}, \qquad \hat{P}(X_u = x_u | SD = h, X_v = x_v) = \frac{\sum_{i \in s_h} I_{x_u, x_v}(x_{iu}, x_{iv})}{\sum_{i \in s_s} I_{x_v}(x_{iv})} \forall\, x_j,\ h.$$

The DJ estimator (3) can consequently be rewritten via the following factorization:

$$\begin{aligned} \hat{P}(x_1, \ldots, x_k) &= \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_{x_1}(x_{i1})}{n_h} \frac{\sum_{i \in s_h} I_{x_1, x_2}(x_{i1}, x_{i2})}{\sum_{i \in s_h}^{H} I_{x_1}(x_{i1})} \cdots \frac{\sum_{i \in s_h} I_{x_1,\ldots,x_k}(x_{i1}, \ldots, x_{ik})}{\sum_{i \in s_h} I_{x_1,\ldots,x_{k-1}}(x_{i1}, \ldots, x_{ik-1})} \\ &= \sum_{h=1}^{H} P(SD = h) \hat{P}(X_1 = x_1 | SD) \prod_{j=2}^{k} \hat{P}(X_j = x_j | SD = h, X_1 = x_1, \ldots, X_{j-1} = x_{j-1}). \end{aligned} \tag{4}$$

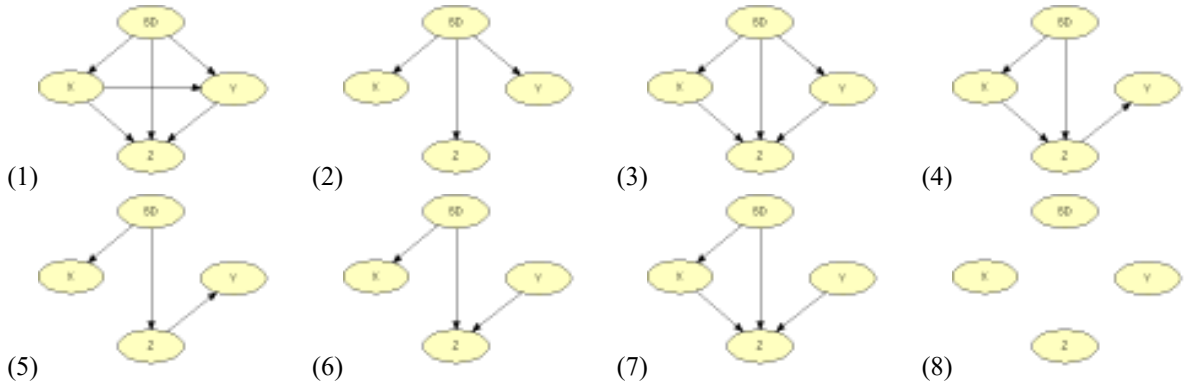


Figure 2: Eight possible BN structures for the nodes (SD, X, Y, Z)

The factorization in (4) corresponds to a particular BN for the variables $(SD, X_1, \ldots, X_k)$: the clique (see Figure 1(1)). As a matter of fact, this is the most complex model for $(SD, X_1, \ldots, X_k)$. When some of these variables are not directly connected (for marginal independence or conditional independence), the complete network is an overparameterized model. Hence, the usual DJ estimator might be less effcient than the one that reproduces the actual dependence model among the variables. For the sake of simplicity, let X, Y and Z be three variables of interest and SD the node representing the survey design. Figure 2 shows 8 different BNs for these variables. The BN (1) shows the already discussed complete network. The other networks are simplifed in the sense that some of the arrows do not appear. Note that it may happen that a variable of interest does not admit SD as a parent. In order to take into account the sample weights also for these variables, the defnition of a BN based estimator of the joint probability distribution would consider 4 different groups of variables (a much simplifed BN based estimator, which disregards sample weights for variables unconnected with SD, is in Ballin *et al*. 2005). The frst two groups are the descendants of SD, while the other two groups are composed of SD non descendants.

Type (a) nodes The nodes of type (a) are all those nodes with SD among their parents. In general, denoting with A the set of labels of the variables of type (a), the estimator of the joint distribution of nodes $\mathbf{X}_A$ is:

$$\hat{P}(\mathbf{X}_A) = \sum_{h=1}^{H} P(SD = h) \prod_{a \in A} P(X_a | pa(X_a)). \tag{5}$$

In Figure 2, networks (1), (2) and (3) have nodes only of type (a), while networks (4), (5), (6) and (7) are such that just X and Z are type (a). For instance the estimator of the joint distribution of (X, Y, Z) for network (1), $\hat{P}_1(x, y, z)$, is defned by (3) or, equivalently, by (4), while the estimators for networks (2) and (3) are respectively:

$$\hat{P}_2(x, y, z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_y(y_i)}{n_h} \frac{\sum_{i \in s_h} I_z(z_i)}{n_h},$$

$$\hat{P}_3(x, y, z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_y(y_i)}{n_h} \frac{\sum_{i \in s_h} I_{xyz}(x_i, y_i, z_i)}{\sum_{i \in s_h} I_{xy}(x_i, y_i)}.$$

The estimator of the (X, Z) distribution for networks (4), (5), (6) and (7) are respectively:

$$\hat{P}^{(4)}(x, z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_{xz}(x_i, z_i)}{\sum_{i \in s_h} I_x(x_i)},$$

$$\hat{P}^{(5)}(x, z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_z(z_i)}{n_h},$$

$$\hat{P}^{(6)}(x, z|y) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_{yz}(y_i, z_i)}{\sum_{i \in s_h} I_y(y_i)},$$

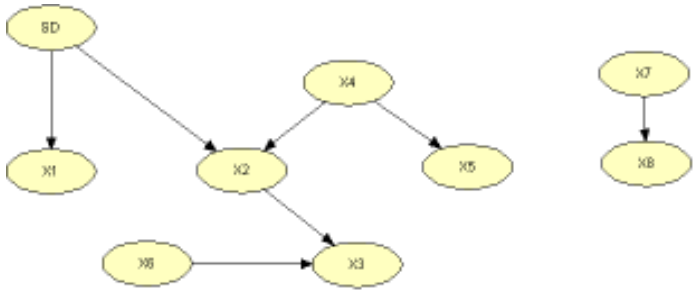


Figure 3: A BN structure for variables $SD, X_1, \ldots, X_8$

$$\hat{P}^{(7)}(x,z|y) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h} \frac{\sum_{i \in s_h} I_{xyz}(x_i, y_i, z_i)}{\sum_{i \in s_h} I_{xy}(x_i, y_i)}.$$

Type (b) nodes A node of type (b) has at least a type (a) ancestor but SD is not one of its parents. For instance, networks (4) and (5) are such that Y is of type (b). In this case, we take as estimators of the joint distribution function for networks (4) and (5) respectively:

$$\hat{P}_4(x,y,z) = \hat{P}^{(4)}(x,z) \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_{yz}(y_i, z_i)}{\sum_{i \in s_h} I_z(z_i)},$$

$$\hat{P}_5(x,y,z) = \hat{P}^{(5)}(x,z) \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_{yz}(y_i, z_i)}{\sum_{i \in s_h} I_z(z_i)}.$$

In order to get these estimators, we implicitly add a fctitious arrow from SD to each type (b) node and estimate its distribution with the survey weights, distinctly from type (a) nodes.

Type (c) nodes This group consists of all those non descendants of SD connected to SD by a non directed path. For instance, Y in networks (6) and (7) is a type (c) node. Also in this case, add a fctitious arrow from SD to the type (c) nodes and estimate their distribution separately from type (a) and (b) nodes. For networks (6) and (7) the estimators are

$$\hat{P}_6(x,y,z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_y(y_i)}{n_h} \hat{P}^{(6)}(x,z|y),$$

$$\hat{P}_7(x,y,z) = \sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_y(y_i)}{n_h} \hat{P}^{(7)}(x,z|y).$$

Type (d) nodes This group consists of all those nodes (or groups of nodes) unconnected with SD, type (a), (b) and (c) nodes. For instance, network (8) consists of 3 different isolated nodes. Again, add a fctitious arrow from SD to each isolated node or group of nodes and estimate them distinctly. The estimator of the joint distribution defned via the BN in network (8) is:

$$\hat{P}_8(x,y,z) = \left[\sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_x(x_i)}{n_h}\right] \left[\sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_y(y_i)}{n_h}\right] \left[\sum_{h=1}^{H} \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{\sum_{i \in s_h} I_z(z_i)}{n_h}\right].$$

While nodes of type (a) form always a single factor that should be marginalized with respect to SD, the other groups may be partitioned in distinct subgroups (as in the example for network (8) in Figure 2). In general these subgroups are separated by type (a) or (b) nodes, while for the nodes in the same subgroup it is possible to fnd a path composed just of nodes of the same subgroup connecting any pair of nodes. An example is presented in Figure 3. Here the nodes are of the following types: X_1 and X_2 are type (a); X_3 is type (b); X_4 and X_5 are frst type (c) subgroup; X_6 is a second type (c) subgroup; X_7 and X_8 are the only type (d) group. According to this partition, the joint distribution function should estimate the following components: $P(X_1, X_2|X_4)$, $P(X_3|X_2, X_6)$, $P(X_4, X_5)$, $P(X_6)$, $P(X_7, X_8)$.

Network	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}_3$	$\hat{P}_4$	$\hat{P}_5$	$\hat{P}_6$	$\hat{P}_7$	$\hat{P}_8$
1	**35.59**	63.84	39.27	59.93	78.49	83.42	58.16	98.89
2	31.39	**18.34**	26.74	49.64	48.09	188.44	194.56	250.07
3	34.76	53.06	**31.47**	135.96	143.86	150.52	133.87	152.25
4	34.63	749.77	53.98	**23.80**	68.12	59.03	44.49	765.77
5	35.99	45.56	31.65	25.39	**23.81**	26.80	32.57	158.67
6	34.52	31.88	29.39	31.30	29.83	**21.17**	26.64	76.06
7	36.80	38.30	31.83	27.06	40.96	42.93	**30.50**	28.35
8	43.27	15.80	37.01	27.78	23.22	23.61	37.19	**15.69**

Table 1: Average of the chi-square distances of the actual and estimated joint distribution of the 8 populations generated according to the 8 networks of Figure 2.

In general, let T, V, and W be the number of these subgroups respectively for type (b), (c) and (d) nodes. Further, let B_t, $t = 1, \ldots, T$, C_v, $v = 1, \ldots, V$ and D_w, $w = 1, \ldots, W$ the set of labels of the variables in each subgroup. Then, the general form of the estimator based on the BN is:

$$\begin{aligned}
\hat{P}(X_1, \ldots, X_k) &= \hat{P}\left(\mathbf{X}_A, \mathbf{X}_{B_1}, \ldots, \mathbf{X}_{B_T}, \mathbf{X}_{C_1}, \ldots, \mathbf{X}_{C_V}, \mathbf{X}_{D_1}, \ldots, \mathbf{X}_{D_W}\right) \\
&= \left[\prod_{w=1}^{W} \hat{P}(\mathbf{X}_{D_w})\right] \left[\prod_{v=1}^{V} \hat{P}(\mathbf{X}_{C_v})\right] \left[\hat{P}(\mathbf{X}_A|\mathbf{X}_{C_1}, \ldots, \mathbf{X}_{C_V})\right] \left[\prod_{t=1}^{T} \hat{P}(\mathbf{X}_{B_t}|\mathbf{X}_A, \mathbf{X}_{C_1}, \ldots, \mathbf{X}_{C_V})\right] \quad (6) \\
&= \left[\prod_{w=1}^{W} \sum_{h=1}^{H} P(SD) \prod_{d \in D_w} \hat{P}(X_d|SD, pa(X_d))\right] \left[\prod_{v=1}^{V} \sum_{h=1}^{H} P(SD) \prod_{c \in C_v} \hat{P}(X_c|SD, pa(X_c))\right] \\
&\quad \left[\sum_{h=1}^{H} P(SD) \prod_{a \in A} \hat{P}(X_a|pa(X_a))\right] \left[\prod_{t=1}^{T} \sum_{h=1}^{H} P(SD) \prod_{b \in B_t} \hat{P}(X_b|SD, pa(X_b))\right].
\end{aligned}$$

As a matter of fact, the distinction among the 4 different types of nodes seems unnecessary. If a node is not directly linked with SD, its (conditional) distribution should be estimated without any marginalization with respect to $P(SD)$. In other words, they should be estimated without weighting the units with their sample weights. However, this approach would be correct when independence holds in the strict sense of its defnition (the joint distribution factorizes in the product of the marginal ones). Although the data generating model fulflls independence in the strict sense, the population generally just fts this model (i.e. if a test had been applied, the hypothesis of independence would have not been rejected). The use of unweighted estimators for type (b), (c) and (d) nodes is illustrated in Ballin *et al.* (2005).
Note that, in any case, the BN based estimator corresponds to a change in the form of the estimator suggested (or in other words *assisted*) by the variable dependence model (the structure of the BN).

2.1 Comparison via a Monte Carlo simulation

Eight populations of 10 000 units with variables X (2 states), Y (3 states) and Z (2 states) have been generated according to the 8 networks in Figure 2. From each population, 500 samples with 1 000 units have been extracted with a simple stratifed sampling design with three strata. Table 1 shows the results of the average chi-square distance between the actual frequency distribution in the population of 10 000 units and the estimated ones. The BN based estimator always wins. It is also robust against mild misspecifcation of the BN. For instance, for the population generated from network (5) the estimator $\hat{P}_4$ is second best. Note also that the usual DJ estimator $\hat{P}_1$ performs almost identically in all the 8 populations. On the contrary, the other BN based estimators can be very ineffcient when the population structure is markedly different from the one that defnes the estimator.

2.2 Relationship between the BN based estimators and the DJ estimator

Which is the relation between the DJ estimator (3) and the BN estimator (6)? Is the BN estimator a linear estimator in the sample weights for at least some marginal distributions? The following proposition details necessary and suffcient conditions for equality between the results of the two estimators.
Proposition 1 - *Let $\hat{P}(\mathbf{X})$ be the BN based estimator of the joint distribution function of $\mathbf{X} = (X_1, \ldots, X_k)$. Let $\mathbf{X}_{sub}$ be a subvector of $\mathbf{X}$. Let $\hat{P}(\mathbf{X}_{sub})$ be the estimator of the joint distribution of $\mathbf{X}_{sub}$ obtained through marginalization. $\hat{P}(\mathbf{X}_{sub})$*

coincides with the one defne d by the DJ estimator if and only if:
(1) $\mathbf{X}_{sub}$ *is composed of variables of the same type and in the same subgroup;*
(2) $\mathbf{X}_{sub}$ *is in a clique with* SD *(after the inclusion of the fctitious arr ows for nodes of type (b), (c) and (d));*
(3) each pair of parents of $\mathbf{X}_{sub}$ *and of its ancestors is joined by an arrow;*
(4) $\mathbf{X}_{sub}$ *admits only ancestors of its type.*
Issue (4) implies that any marginal and joint distribution involving type (b) nodes can never be estimated as DJ estimators. As an example, the BN in Figure 3 implies that X_1, X_2, (X_4, X_5), X_6 and (X_7, X_8) are the maximal tables that can be estimated according to the DJ.

2.3 BN as a tool for incorporating further information
BNs can be updated when new information is available (*informative shock*). Information is in terms of a new frequency distribution for one or more of the variables of interest gained from an archive or a new survey. The relationship among the variables of a BN (i.e. the arrows) are the highway for the propagation of this kind of information. For the sake of simplicity, let the BN be composed of just two nodes, X_1 and X_2, joined by the arrow $X_1 -> X_2$. Hence, the BN is composed of the following probability distributions: $P(X_1 = x_1)$, $P(X_2 = x_2|X_1 = x_1)$. Let the marginal probability distribution for X_2 be changed in: $P^*(X_2 = x_2)$. In order for the network to incorporate the new distribution $P^*(X_2 = x_2)$ leaving unchanged the relationship between the variables, i.e. the conditional distribution of X_2 given X_1, it is necessary to modify the marginal distribution of X_1:

$$P^*(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1|X_2 = x_2)P^*(X_2 = x_2) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)\frac{P^*(X_2 = x_2)}{P(X_2 = x_2)}.$$

In other words, the old joint distribution $P(X_1 = x_1, X_2 = x_2)$ is updated via the ratio of the new and old marginal distributions of X_2:

$$\frac{P^*(X_2 = x_2)}{P(X_2 = x_2)}. \tag{7}$$

What explained for two variables can be generalized for general situations when new information updates more than one variable distribution. To this purpose, different effcient algorithms based on the concept of *junction trees* (see Jensen, 1996) have been defned.

2.4 Poststratifcation
What described in the previous paragraph for a general BN can be easily applied for the traditional poststratifcation procedure in fnite survey sampling. Let us consider the usual DJ estimator (3) or, equivalently the BN based estimator corresponding to the clique (4). For the sake of simplicity let the informative shock be relative to just variable X_1 in the following situation: a sample $\mathcal{S}$ is drawn according to a design which is not stratifed with respect to X_1, and we have an informative shock on the X_1 frequency distribution, say N^*_{1q}, $q = 1, \ldots, Q$. This informative shock can be used in order to poststratify the sample with respect to X_1. The old sample weights ω_i are consequently changed into:

$$\omega^*_i = \omega_i \frac{N^*_{1q}}{\sum_i \omega_i I_{x_{1i}}(q)} = \omega_i \frac{N^*_{1q}}{\hat{N}_{1q}}, \qquad i:\ I_{x_{1i}}(q) = 1,\ q = 1, \ldots, Q \tag{8}$$

where $\hat{N}_{1q}$ are the frequency estimates computed on the old survey weights. This operation is quite similar to the one in (7). In fact, the change is in the node SD, which modifes into a new node SD^* with the following characteristics: (i) SD^* categories are given by the Cartesian product of the SD and X_1 categories i.e. (h, q), $h = 1, \ldots, H$, $q = 1, \ldots, Q$; (ii) the units in the same category (h, q) have the same weight, $\omega^*_{(h,q)}$. Again, Bayes theorem allows the computation of the probability distribution of SD given X_1:

$$P(SD = h|X_1 = q) = \frac{P(SD = h)P(X_1 = q|SD = h)}{\sum_{h=1}^{H} P(SD = h)P(X_1 = q|SD = h)},\ q = 1, \ldots, Q; h = 1, \ldots, H.$$

Leaving unchanged the previous distribution, i.e. the statistical relationship between SD and X_1 according to the initial survey design, poststratifcation with respect to the new distribution of X_1, $N^*_1(q)$, $q = 1, \ldots, Q$, or better to the relative frequency

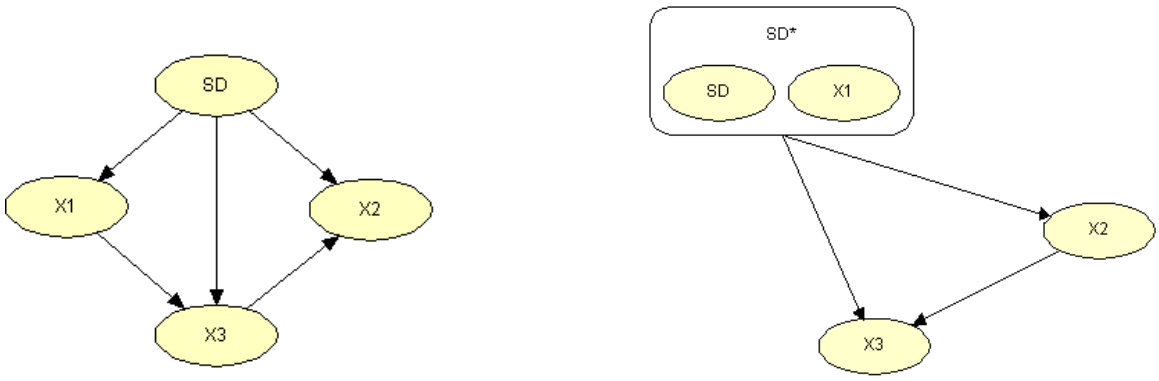


Figure 4: Transformation of a BN after poststratifcation.

distribution: $P_1^*(q) = N_{1q}^*/N$, $q = 1, \ldots, Q$, corresponds to consider this new joint distribution:

$$\begin{aligned} P(SD^* = (h,q)) &= P(SD = h, X_1 = q) = P(SD = h|X_1 = q)P_1^*(q) = \\ &= \frac{P(SD = h)P(X_1 = q|SD = h)}{\sum_{h=1}^{H} P(SD = q)P(X_1 = q|SD = h)} P_1^*(q) = \\ &= \frac{n_h \omega_{(h)}}{\sum_{h=1}^{H} n_h \omega_{(h)}} \frac{n_{hq}}{n_h} \frac{P_1^*(q)}{\hat{P}_1(q)}, \qquad q = 1, \ldots, Q; h = 1, \ldots, H. \end{aligned} \tag{9}$$

The new weight $\omega_{(h,q)}^*$ must be constant for all the units in the same SD^* category, of size n_{hq}. Hence:

$$\omega_{(h,q)}^* = \frac{\sum_{h=1}^{H} n_h \omega_{(h)}}{n_{hq}} P(SD^* = (h,q)) = \omega_{(h)} \frac{P_1^*(q)}{\hat{P}_1(q)} = \omega_{(h)} \frac{N_1^*(q)}{\hat{N}_1(q)} \frac{\hat{N}}{N} \tag{10}$$

where $\hat{N} = \sum_{i=1}^{N} \omega_i$.

This procedure remains unchanged in case the variables are all of the same type in the same subgroup. For example consider Figure 4. The network on the left contains only type (a) nodes. Poststratifcation with respect to X_1 produces the network on the right where the design variable SD^* is now the clique (SD, X_1).

As shown before, the BN representation allows to propose many estimators according to the different types of variables. Therefore some issues need further study.

(1) Which is the role of different network structures for SD^*.

(2) How to defne poststratif cation when variables of different type defne the BN.

(3) How to poststratify when joint information on variables of different types is available.

(4) What connection there is with ratio raking estimators.

(5) How to generalize this procedure to the case of integration of two or more surveys, as in Ballin and Vicard (2001).

References

Ballin M, Vicard P (2001). A proposal for the use of graphical representation in offcial statistics, *Proceedings SCO2001*, 24-26 September 2001, Bressanone, Italy.

Ballin M, Scanu M, Vicard P (2005). Model assisted approaches to complex survey sampling from fnite populations using Bayesian networks: a tool for integration of different sources. *Proceedings of XXII Statistics Canada International Methodology Symposium*, Ottawa, October 25–28, 2005.

Cowell R G, Dawid A P, Lauritzen S L, Spiegelhalter D J (1999). *Probabilistic Networks and Expert Systems*, Springer.

Jensen F V (1996). *An Introduction to Bayesian Networks*, Springer.

Neapolitan, R E (2004). *Learning Bayesian Networks*, Prentice Hall.

Särndal C E, Swensson B, Wretman J (1992). *Model Assisted Survey Sampling*, Springer.