Journal of Official Statistics, Vol. 26, No. 4, 2010, pp. 673–694

Statistical Model of the 2001 Czech Census for Interactive Presentation

Jiří Grim[1], *Jan Hora*[2], *Pavel Boček*[1], *Petr Somol*[1], *and Pavel Pudil*[3]

This article describes the application of a recently developed method of interactive statistical database presentation to the 2001 Czech Census. The method is based on estimating the multivariate probability distribution of the original microdata, which are supposed to be discrete or discretized continuous. The estimated statistical model in the form of a distribution mixture of product components can be used as a knowledge base of a probabilistic expert system. By means of the probabilistic inference mechanism we can derive conditional distributions of variables for each subpopulation interactively without any further access to the source database. The conditional distributions (histograms) describing the properties of subpopulations represent the basic form of user information. The statistical model does not contain the original data and therefore can be distributed without any confidentiality concerns. The accuracy achievable by the statistical model is comparable with that of the anonymized subsets of microdata.

Key words: Interactive statistical model; census data presentation; distribution mixtures; data modeling; EM algorithm; incomplete data; data reproduction accuracy; data mining.

1. Introduction

In our modern networked society there is an increasing demand for dissemination and sharing of statistical information. To meet the expectations of users, statistical agencies release two major forms of statistical data: traditional tabular data and the sets of individual respondent records called microdata. The advantage of releasing microdata instead of specific precomputed tables and statistics follows from the increased flexibility and availability of information for the users. With appropriate microdata, the users may examine unusual hypotheses and find new issues beyond the usual scope of data providers.

In any case, the fundamental obligation of data providers is to protect the privacy of respondents. For this reason, explicit identifiers such as names, addresses and phone numbers are commonly removed. However, anonymous respondents may by reidentified by combining other data such as birth date, sex, and ZIP code which uniquely pertain to specific individuals. Different statistical disclosure control (SDC) methods have been proposed to protect the confidentiality of data. With tabular data a disclosure can occur

[1] Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic. Email: grim@utia.cas.cz
[2] Faculty of Nuclear Science and Physical Engineering, Czech Technical University, Czech Republic
[3] Faculty of Management, Prague University of Economics, Czech Republic

Acknowledgment: This study was partially supported by the project GAČR No. 102/07/1594 of Czech Grant Agency and by the projects 2C06019 ZIMOLEZ and MSMT 1M0572 DAR.

if a cell corresponds to a very small group of respondents. This problem can be eliminated by suppressing cells, aggregating values, removing sensitive variables or by other techniques. In the case of microdata, easily identifiable quantitative variables may be transformed to discrete intervals, and sensitive qualitative variables may be combined to produce more general categories. Rare data can be suppressed, swapped, modified or simulated. Obviously, disclosure limitation procedures are connected with some information loss.

There is extensive literature on SDC techniques available (for the most frequent references see, e.g., Dalenius 1977; Bethlehem et al. 1990; Winkler 1998; Fienberg 1994; Fienberg et al. 1998; Willenborg and de Waal 2001), but the underlying problem is still to be considered open. Let us recall that a disclosure may be inferential (Duncan and Lambert 1989) without any actual reidentification of a record and, in a special context, even a modified erroneous value may be harmful for the reidentified respondent. Much extra work is needed to produce safe and analytically valid public-use files though there is always a remaining disclosure risk. Thus, many statistical agencies release microdata for research purposes only, usually under special licence agreements and through secure data archives. In general, the nondisclosure policy becomes a serious limitation on information dissemination.

In recent years we have developed an alternative approach to the presentation of survey results based on interactive statistical models (Grim 1992; Grim and Boček 1996; Grim et al. 2001; Grim et al. 2004a, b). We estimate the joint probability distribution of the original discrete microdata in the form of a multivariate distribution mixture with product components using the EM algorithm (Dempster et al. 1977). We assume the variables to be discrete (ordinal or categorical, qualitative). Continuous variables have to be discretized by introducing intervals. The estimated product mixture can be used as a knowledge base of a probabilistic expert system PES (Grim 1990; Grim 1994). We can thus derive the statistical information from the mixture model without any further access to the original database. The statistical model provides flexibility and comfort of information analysis which in some respects is comparable to, or even better than, microdata subsets.

The mixture model describes the statistical properties of microdata in terms of univariate component-specific probability distributions. By its nature, any information derived from the mixture model is a (conditional) probability. Thus, any model-estimated cell counts are biased by uncertainty, which considerably increases at low probability values. Even if the estimated cell count approaches unity, there is no guarantee that a corresponding record uniquely exists in the original database. The final software product does not contain any original or synthetic microdata, and not even the model parameters are directly available for users. Since there is no possibility of identifying any concrete respondent from regular univariate distributions, the model-based interactive software can be distributed without any confidentiality concerns.

A weak point of the method is the model accuracy. The interactive software does not provide exact values and it is not suitable for reproducing the statistical properties of continuous variables or of discrete variables with a great number of possible values (small area identifiers, detailed age groups). Nevertheless, we assume that the limited applicability of the method is well counterbalanced by the possibility of unrestricted distribution of the final software product which could improve the information offer of statistical offices to the public. To the best of our knowledge, in the recent literature there are no similar results on statistical data models based on distribution mixtures.

In this article we describe an application of the proposed method to the individual microdata records from the 2001 Czech Census. The statistical model has been computed within the framework of a special cooperation project between the Czech Statistical Office, Prague University of Economics and the Institute of Information Theory and Automation. The aim of the project is to verify the applicability of the interactive statistical model to the next Czech Census in 2011. To illustrate a general possibility of information fusion from different sources, we have combined the databases of persons and households that were originally treated separately.

The resulting source database contained 10,230,060 records, with about 1.5 million incomplete records including nearly three million nonresponse (missing) values. As the primary purpose of the project has been to demonstrate the accuracy of the method in the case of ideal complete data, we decided first to estimate the model parameters from the incomplete records, and then to use the resulting distribution mixture to estimate and substitute the missing values. The final statistical model has been computed from the "ideal" set of complete microdata. The accuracy of the final model has been verified by comparing the model probabilities with the relative frequencies of all statistically relevant combinations of responses in the completed database. We have established that the accuracy of model probabilities is comparable to that of the relative frequencies computed from a randomly chosen one-million subset of the original microdata (without anonymization). The preliminary version of the final interactive software product can be downloaded at our webpage http://ro.utia.cas.cz/dem.html (Data Mining: Interactive Presentation of Census Results by Probabilistic Models).

The article is organized as follows: in Section 2 we describe the choice of variables for the statistical model, the EM algorithm and its properties. Section 3 deals with the problem of missing data and in Section 4 we evaluate the accuracy of the estimated mixture. In Section 5 we discuss some tools of information analysis and in the concluding section we summarize advantages and different application aspects of the proposed method.

2. Statistical Model of Census Data

The primary purpose of the considered statistical model is to reproduce the statistical relationships within a given finite set of discrete variables as exactly as possible. The number of variables and number of their values should be kept in reasonable bounds because of the well-known trade-off between the complexity of the estimated probability distribution and its accuracy. For the sake of estimating the statistical model of the 2001 Czech Census we have chosen 24 categorial variables (questions) as listed in Table 1. We have applied less detailed coding of some variables (regional localization, age intervals) to decrease the formal complexity of the model. Simultaneously we have also omitted unambiguous variables, which are less informative and unproductive in combination with other variables.

To illustrate a general possibility of information fusion from different sources we have combined two originally separate databases of individuals and households. In particular, the first ten variables from the database of individuals have been merged with fourteen variables of the corresponding households. Note that in the resulting database the household-related response frequency has a different meaning, namely the number

Table 1. List of questions included in the statistical model of the 2001 Czech Census. The third column contains the number of possible responses, the percentage of missing values (nonresponse) is given in the fourth column. There are 1,524,240 incomplete records, the total number of nonresponses is 2,933,427. Uncertainty of variables in % of maximum Shannon entropy is given in the last column

	Text of question (name of variable)	Number of values	Nonresponse in %	Shannon entropy in %
1.	Region of residence	14	0.00	96.88
2.	Type of residence	3	0.00	32.92
3.	Economic activity	10	0.80	67.80
4.	Birthplace (relatively)	6	1.95	74.65
5.	Religion	6	0.00	60.57
6.	Occupation type	14	3.89	68.33
7.	Sex	2	0.00	99.95
8.	Marital status	4	0.55	81.01
9.	Education	14	1.11	78.04
10.	Age	9	0.03	96.09
11.	Category of flat	5	0.53	27.81
12.	Bathroom	5	0.59	14.02
13.	Size of flat	7	0.64	80.62
14.	Internet and PC	4	2.85	49.11
15.	Legal relation to flat	9	0.39	72.43
16.	Gas supply	3	0.78	64.54
17.	Number of rooms over 8 m^2	7	0.64	80.57
18.	Number of cars in household	4	3.39	71.32
19.	Number of persons in flat	6	0.00	93.79
20.	Vacational property	6	7.45	42.10
21.	Telephone in flat	5	1.80	80.88
22.	Water supply	4	0.35	8.02
23.	Type of heating	6	0.53	74.81
24.	Toilet	6	0.50	16.73

of respondents living in the respective households. Thus, instead of the properties of flats, we may analyze the housing conditions of respondents.

For every respondent we have a record of 24 variables. The third column in Table 1 contains the number of possible responses for the respective questions and the fourth column contains the frequency of missing values as a percentage. The total number of nonresponses is 2,933,427. Uncertainty of variables expressed in percentage of maximum Shannon entropy is given in the last column. The highest uncertainty comes with Question 7 (response "sex" implies two nearly equal response frequencies). In contrast, Question 22 (water supply) implies the most unambiguous response.

Formally, we consider the source database to be a set of independent and identically distributed observations of a random vector of 24 discrete finite valued random variables:

$$v = (v_1, v_2, \ldots, v_{24}) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_{24}. \tag{1}$$

We assume that the unknown multivariate discrete probability distribution $P*(x)$ of the random vector v can be approximated by a finite distribution mixture of product components:

$$P(x) = \sum_{m=1}^{M} w_m F(x|m), \quad F(x|m) = \prod_{n=1}^{24} p_n(x_n|m), \quad x \in \mathcal{X}, \quad \sum_{m=1}^{M} w_m = 1. \tag{2}$$

Here $w_m > 0$ is the prior weight of the m-th mixture component, $p_n(x_n|m)$ are the conditional (component specific) univariate distributions of the variables v_n, and M is the number of components. Note that the product components do not imply that the involved variables are independent. In this sense the mixture model (2) is not restrictive. It is easily verified (cf. Grim and Boček 1996) that by increasing the number of components we can describe any discrete probability distribution in the form (2).

The standard way to estimate the parameters of the distribution mixture (2) is to use the EM algorithm (Schlesinger 1968; Dempster et al. 1977; Grim 1982; Grim 1992; Grim and Boček 1996; Grim et al. 2001; Grim et al. 2004a, b). In particular, let S be a set of K data vectors as obtained, e.g., in census:

$$S = \{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}, \quad x^{(k)} \in \mathcal{X}, \ (K = |S|). \tag{3}$$

To estimate the unknown mixture parameters we maximize the log-likelihood function

$$L = \frac{1}{|S|}\sum_{x \in S} \log P(x) = \frac{1}{|S|}\sum_{x \in S} \log\left[\sum_{m=1}^{M} w_m F(x|m)\right] \tag{4}$$

by means of the following EM iteration equations $(m = 1, 2, \ldots, M, n = 1, 2, \ldots, 24)$:

$$q(m|x) = \frac{w_m \prod_{n=1}^{24} p_n(x_n|m)}{\sum_{j=1}^{M} w_j \prod_{n=1}^{24} p_n(x_n|j)}, \quad w'_m = \frac{1}{|S|}\sum_{x \in S} q(m|x), \ x \in S, \tag{5}$$

$$p'_n(\xi|m) = \frac{1}{\sum_{x \in S} q(m|x)} \sum_{x \in S} \delta(\xi, x_n) q(m|x), \quad \xi \in \mathcal{X}_n. \tag{6}$$

Here the apostrophe denotes the new parameter values and $\delta(\xi, x_n)$ is the delta-function in the usual sense ($\delta(\xi, x_n) = 1$ for $\xi = x_n$ and $\delta(\xi, x_n) = 0$ for $\xi \neq x_n$).

The EM algorithm monotonously converges to a local or global maximum or to a saddle point of the log-likelihood function L in the sense that the corresponding sequence of values $\{L^{(t)}\}_{t=0}^{\infty}$ is nondecreasing (cf. Dempster et al. 1977; Grim 1982). The existence of local maxima makes the procedure starting-point dependent. In this respect a well-known difficulty is to specify the number of mixture components and to choose the initial parameter values (cf., e.g., McLachlan and Peel 2000). However, the problem becomes less relevant in high-dimensional spaces and with increasing number of components since the values of different local maxima are similar and therefore the quality of the corresponding mixture estimates is comparable. For the same reason the mixture parameters have been initialized randomly in all of our experiments.

As can be expected, the accuracy of the model increases with the model complexity, on the other hand, the number of components is the main limiting feature from a computational point of view. For this reason, in the following, the choice of the number of mixture components M is mainly influenced by practical hardware-specific considerations.

As the main purpose of the mixture model (2) is to reproduce the statistical properties of the original data, it is important that, in each iteration of the EM algorithm, the univariate marginal distributions of the estimated mixture are identical with the global marginal frequencies of the data. In particular, by using Equations (5) and (6), we can write

$$\begin{aligned} P'_n(\xi) &= \sum_{m=1}^{M} w'_m p'_n(\xi|m) = \sum_{m=1}^{M} \frac{1}{|S|}\sum_{x\in S} q(m|x) p'_n(\xi|m) = \\ &= \frac{1}{|S|}\sum_{x\in S}\delta(\xi,x_n)\sum_{m=1}^{M} q(m|x) = \frac{1}{|S|}\sum_{x\in S}\delta(\xi,x_n), \quad \xi\in\mathcal{X}_n,\ n=1,2,\ldots,24. \end{aligned} \tag{7}$$

Recall that any marginal distribution of the mixture (2) is easily obtained by ignoring superfluous terms in the products. In view of this property the discrete distribution mixture (2) is directly applicable as a knowledge base of the Probabilistic Expert System (PES) (cf. Grim 1994; Grim and Boček 1996). In this way the inference mechanism of PES (cf. Appendix I) derives the statistical properties of different subpopulations directly from the estimated model without any access to the original microdata and without using any synthetic data.

In particular, considering a given input subvector

$$x_C = (x_{i1}, x_{i2}, \ldots, x_{ik}) \in \mathcal{X}_C, \quad C = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, 24\},$$

and an output variable x_n, $(n \notin C)$, we can directly write equations for the related marginal

$$P_C(x_C) = \sum_{m=1}^{M} w_m F_C(x_C|m), \quad F_C(x_C|m) = \prod_{i\in C} p_i(x_i|m), \quad x_C \in \mathcal{X}_C, \tag{8}$$

and for the corresponding conditional distribution

$$P_{n|C}(x_n|x_C) = \frac{P_{n,C}(x_n, x_C)}{P_C(x_C)} = \sum_{m=1}^{M} W_m(x_C) p_n(x_n|m), \quad (P_C(x_C) > 0). \tag{9}$$

Here $W_m(x_C)$ are the conditional component weights for the given subvector $x_C \in \mathcal{X}_C$:

$$W_m(x_C) = \frac{w_m F_C(x_C|m)}{\sum_{j=1}^{M} w_j F_C(x_C|j)} \tag{10}$$

Note that the conditional distributions $P_{n|C}(x_n|x_C)$ (conditional histograms) describe the statistical properties of the subpopulation specified by the subvector x_C in terms of all variables x_n not included in x_C. For a given input x_C, Equation (9) is applicable to different variables $n \notin C$ with identical weights $W_m(x_C)$. Thus, for any fixed subvector x_C, we obtain a set of histograms which characterize the corresponding subpopulation. We can efficiently store extensive lists of subpopulations in terms of the defining subvectors. In this way different subpopulations can be quickly compared and characterized in terms of conditional histograms (cf. Figure 1), e.g., by the most apparent differences from the whole population. From the point of view of a user the conditional histograms describing the properties of different subpopulations represent a basic form of statistical inference.

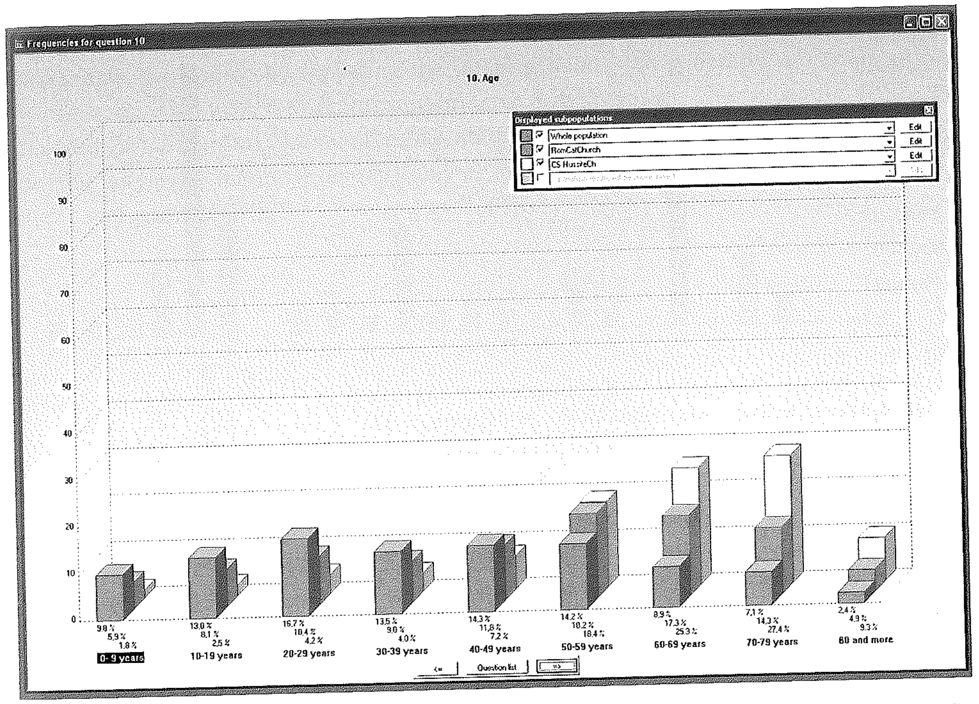

Fig. 1. Comparison of age distribution in three different subpopulations (cf. http://ro.utia.cas.cz/dem.html)

In addition, the analytical simplicity of the statistical model suggests some new possibilities of information analysis (cf. Section 4).

3. Missing Data

A typical feature of census data is the presence of incomplete records. The census database considered in this article (cf. Table 1) included 1,524,240 incomplete records containing up to eighteen missing values. The distribution of nonresponse according to variables is given in Figure 2. Figure 3 displays the distribution of incomplete records by the number of missing values. The total number of missing values in our database was 2,933,427.

The problem of missing data is traditionally an important area of mathematical statistics, because most statistical methods cannot be applied to incomplete data. One can see that by simply omitting the incomplete records we would lose about 15% of the records in our database. Similarly, only five questions would remain should we ignore incomplete variables. Since the missing values are denoted as nonresponse in all records, they could always be treated as a specific additional response. However, in most cases, the information value of nonresponse is limited because of its latent dependency on the context (there are multiple modes of nonresponse, e.g., don't know and refuse). In this sense the additional value would cause superfluous increase of relationship complexity; it means we would force the model to describe many meaningless statistical relations.

An important feature of estimating product mixtures is the possibility of modifying the EM algorithm to be directly applicable to incomplete data. In this sense there is no

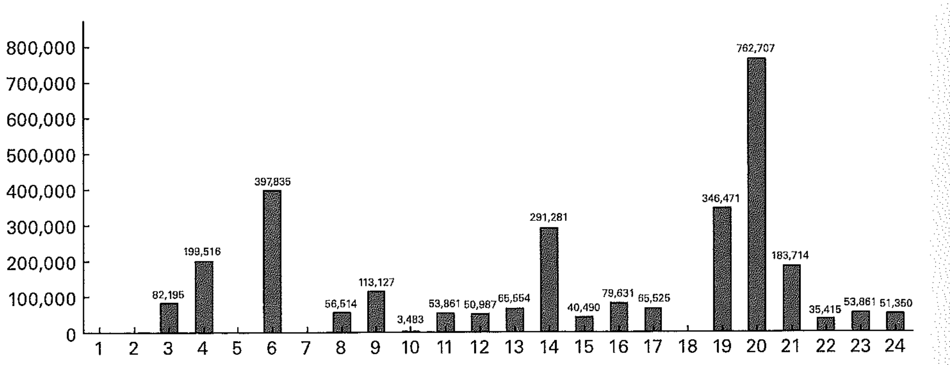

Fig. 2. Nonresponse frequency for individual questions. The number of incomplete records is 1,524,240, the total number of missing values is 2,933,427

necessity to substitute for the missing values; we estimate the mixture parameters from the available data only. The type of missing values is nearly irrelevant since the estimated model can utilize all statistical information in the data. In particular, denoting by $\mathcal{N}(x)$, the subset of indices of the available variables of x, and by $S_n \subset S$, the subset of vectors with the available variable x_n:

$$\mathcal{N}(x) = \{n : x_n \,\text{available in}\, x\}, \quad S_n = \{x \in S : n \in \mathcal{N}(x)\}, \tag{11}$$

we can write the modified EM iteration equations in the form ($m = 1, 2, \ldots, M$, $n = 1, 2, \ldots, 24$, $x \in S$):

$$q(m|x) = \frac{w_m \prod_{n \in \mathcal{N}(x)} p_n(x_n|m)}{\sum_{j=1}^{M} w_j \prod_{n \in \mathcal{N}(x)} p_n(x_n|j)}, \quad w'_m = \frac{1}{|S|}\sum_{x \in S} q(m|x), \tag{12}$$

$$p'_n(\xi|m) = \frac{1}{\sum_{x \in S_n} q(m|x)} \sum_{x \in S_n} \delta(\xi, x_n) q(m|x), \quad \xi \in \mathcal{X}_n. \tag{13}$$

Roughly speaking, we calculate the values $q(m|x)$ *a* $p'_n(\xi|m)$ in Equations (12) and (13) only for the variables currently available in x.

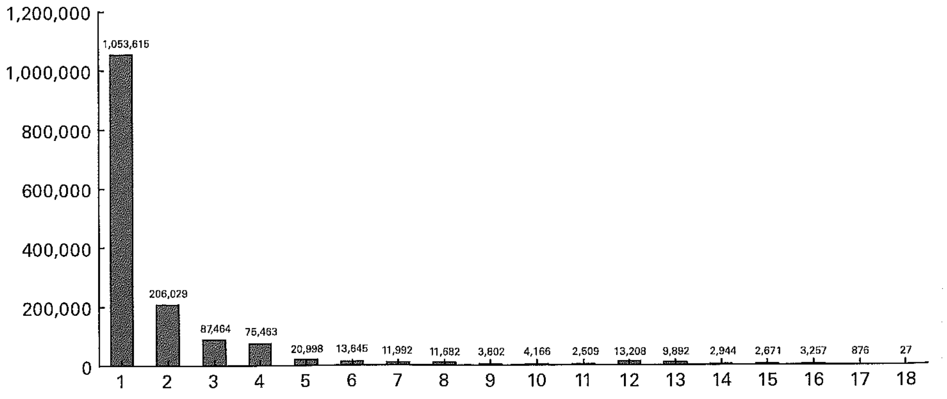

Fig. 3. Distribution of incomplete records according to the number of missing values

From the theoretical point of view we would prefer the direct estimation from incomplete data, since by replacing the missing values we generally use some typical values, and in this way, the natural variability of data decreases. Unfortunately, unlike the standard EM algorithm, the modified iteration Equations (12) and (13) do not guarantee that the univariate marginal distributions of the estimated mixture are identical with the global marginal frequencies of the data (cf. (7)):

$$\begin{aligned} P'_n(\xi) &= \sum_{m=1}^{M} w'_m p'_n(\xi|m) = \sum_{m=1}^{M} \frac{1}{|S|} \sum_{x \in S} q(m|x) p'_n(\xi|m) = \\ &= \frac{1}{|S|} \sum_{m=1}^{M} \frac{\sum_{x \in S} q(m|x)}{\sum_{x \in S_n} q(m|x)} \sum_{x \in S_n} \delta(\xi, x_n) q(m|x) \neq \frac{1}{|S|} \sum_{x \in S} \delta(\xi, x_n), \quad \xi \in \mathcal{X}_n, \end{aligned} \tag{14}$$

It appears as if the mixture model estimated from incomplete data is biased by a considerable error already at the level of unconditional marginals. In additional experiments we have established that a model obtained from incomplete data is approximately twice as inaccurate as the comparable model computed from complete data.

Let us recall that the main purpose of our project has been to verify the possibility of reproducing the statistical properties of a large set of microdata and therefore the model accuracy is of primary importance. For this reason we decided to solve the estimation problem in two steps. First we estimated the distribution mixture (2) from incomplete data by means of the modified EM algorithm (12) and (13). The resulting mixture (M = 10,000) has been used to replace missing values by estimates. In particular, we have replaced each missing value by the response x_n, which is most probable in the sense of the conditional distribution $p_n(x_n|x_c)$ (cf. (9)):

$$x_n = \arg\max_{\xi \in \mathcal{X}_n} \{p_n(\xi|x_C)\} \tag{15}$$

In other words, we have replaced each nonresponse by the value $x_n \in \mathcal{X}_n$, which is the most probable response with respect to the known part x_C of the record. In the second step we have used the completed database to estimate the final distribution mixture.

Generally, the computing time of the EM algorithm is proportional to the model complexity; therefore, in the case of the given large database, the number of mixture components is the most relevant practical limitation. We have chosen M = 15,000 components for the final experiment. Again, we generated the initial parameters randomly and the computation has been stopped at the point of sufficient convergence after approximately thirty iterations. It took about eight hours per iteration (on a standard PC), which corresponds to a total computing time of about ten days. As a stopping rule for EM iterations we have used the threshold $\varepsilon = 10^{-4}$ for the relative increment of the log-likelihood criterion (4). Nevertheless, in the final stages of convergence the model accuracy does not change very much and the algorithm can be stopped manually.

It is obvious that the imputation of missing values may affect the final model accuracy. There is no direct possibility of verifying if the replacement of missing values has been done correctly but we can simulate an analogous situation by estimating the known values (cf. Williams 2005). In particular, for each variable separately, we have randomly chosen 10^5 records for which the value of the tested variable was available. Then for each record we have computed the corresponding estimate of this value and compared it with the true original. The results of the imputation test are summarized in Table 2, which provides additional information about the accuracy of the final statistical model ($M = 15{,}000$). The third column contains the number of nonresponses and in the fourth column we list the percentage of correctly estimated values. The number in parentheses corresponds to the trivial global imputation of the most frequent response. As one would expect, the imputation accuracy is variable-dependent. In some cases the success of global imputation of the most frequent value is comparable with the statistical model (Nos. 2, 4, 5, 12, 14, 20, 22) but the improvement achieved by using the maximum-likelihood estimate is often considerable (Nos. 1, 3, 8, 9, 10, 13, 15, 17, 19, 23). On average, 73% of missing values would be correctly identified by the maximum-likelihood estimates. The last column contains the number of nonresponses from the third column that are expected to be replaced correctly.

Table 2. Accuracy of the estimation of missing values. The third column contains the number of nonresponses. In the fourth column we list the percentage of correctly estimated responses. The numbers in parentheses correspond to the trivial global imputation of the most frequent response. The last column lists expected numbers of correctly replaced nonresponse from the column

	Text of question (name of variable)	Number of nonresponses	Successful imputation in %	Successful imputation
1.	Region of residence	0	27.49 (12.41)	0
2.	Type of residence	0	90.35 (89.48)	0
3.	Economic activity	82,195	88.02 (44.08)	72,348
4.	Birth place (relatively)	199,516	56.36 (53.52)	112,447
5.	Religion	0	66.27 (59.04)	0
6.	Occupation type	397,835	67.64 (50.62)	269,096
7.	Sex	0	67.91 (51.30)	0
8.	Marital status	56,514	82.91 (46.63)	46,856
9.	Education	113,127	48.36 (19.29)	54,708
10.	Age	3,483	59.22 (16.71)	2,063
11.	Category of flat	53,861	97.48 (89.37)	52,504
12.	Bathroom	50,987	98.90 (95.91)	50,426
13.	Size of flat	65,554	63.22 (38.48)	41,443
14.	Internet and PC	291,281	81.12 (79.15)	236,287
15.	Legal relation to flat	40,490	63.49 (39.70)	25,707
16.	Gas supply	79,631	75.94 (63.84)	60,472
17.	Number of rooms over 8 m^2	65,525	63.48 (38.76)	41,595
18.	Number of cars in household	346,471	66.97 (51.77)	232,032
19.	Number of persons in flat	0	49.48 (29.27)	0
20.	Vacational property	762,707	80.39 (78.11)	613,140
21.	Telephone in flat	183,714	57.36 (43.93)	105,378
22.	Water supply	35,415	99.39 (98.08)	35,199
23.	Type of heating	53,861	76.90 (41.45)	41,419
24.	Toilet	51,350	97.98 (94.32)	50,313
	Total	2,933,427	73.06 (61.35)	2,143,326

4. Accuracy of the Statistical Model

Let us recall that the primary purpose of the estimated model is to reproduce the statistical properties of the modeled data. In the domain of statistical surveys, we usually specify the subpopulations by combining responses. Therefore the statistical model should reproduce the empirical frequencies of different properties as precisely as possible. In particular, in order to verify the model accuracy, we compare the empirical frequencies of different combinations of responses with the estimates derived from the statistical model. The accuracy of the final model has been verified by the underlying completed data set, to demonstrate the performance of the method in the "ideal situation" without any unpredictable influence of missing data.

Considering an elementary property defined by a subvector of responses x_C, we denote

$$S(x_C) = \{y \in S : y_C = x_C\}, \quad N(x_C) = |S(x_C)|, \quad x_C = (x_{i1}, \ldots, x_{ik}) \in \mathcal{X}_C \tag{16}$$

where $S(x_C)$ is the subset of respondents (a subpopulation) with the property x_C and $N(x_C)$ is the (empirical) frequency of the property x_C in the census population S. Obviously, the frequency $N(x_C)$ can be estimated from the statistical model (2) as the product of the probability $P(x_C)$ and the population size $|S|$:

$$\hat{N}(x_C) = |S|P(x_C), \quad P(x_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{k} p_{i_j}(x_{i_j}|m) \tag{17}$$

It appears that, ideally, we should compare the estimated frequency $\hat{N}(x_C)$ with the empirical value $N(x_C)$ for all possible elementary combinations of values x_C. However, there are two important limitations.

Recall first that we are not interested in reproducing small frequencies. On the contrary, the decreasing accuracy of the model at low probabilities is an important confidentiality-protecting property. We decided for this reason to evaluate the accuracy of the estimates $\hat{N}(x_C)$ only for the empirical frequencies $N(x_C)$ greater than a suitably chosen threshold N_ε. In order to specify the threshold frequency N_ε, we confine ourselves only to "statistically relevant" properties x_C, the frequency of which may differ from the assumed "true" unknown frequency $N*(x_C)$ by less than $\varepsilon = 5\%$ (cf. Appendix II). In particular, if we confine ourselves to the properties x_C satisfying the inequality $N(x_C) > 1{,}612$ (i.e., $N_\varepsilon = N_{0.05} = 1{,}612$), then, according to the central limit theorem of probability theory, the empirical frequency $N(x_C)$ of the property x_C in the population S may differ from the unknown "true" frequency $N*(x_C)$ by less than 5% (at the confidence level 0.95).

The second limitation has a computational origin. The number of all properties x_C specified by all possible combinations of responses is too high and the evaluation would be too time-consuming. For this reason we decided to verify the model accuracy by considering combinations of a maximum of five responses. As a result we obtained a list $\mathcal{A}_5$ of about 26 million "statistically relevant" properties x_C along with the corresponding empirical frequencies

$$\mathcal{A}_5 = \{x_C = (x_{i1}, \ldots, x_{i5}) : N(x_C) > 1{,}612\}, \quad |\mathcal{A}_5| = 26{,}425{,}727. \tag{18}$$

A natural way to measure the accuracy of the statistical model (2) is to compute the mean absolute error E_a of the estimated frequencies $\hat{N}(x_C)$ for the properties $x_C \in \mathcal{A}_5$:

$$E_a = \frac{1}{|\mathcal{A}_5|}\sum_{x_C \in \mathcal{A}_5} |P(x_C)|S| - N(x_C)|, \quad P(x_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{5} p_{i_j}(x_{i_j}|m) \tag{19}$$

where $P(x_C)$ is the probability of the combination x_C computed by means of the mixture model (2). However, as can be seen, the criterion E_a does not differentiate between errors of large and small estimates. Therefore we have introduced the following mean relative error criterion

$$E_r = \frac{100}{|\mathcal{A}_5|}\sum_{x_C \in \mathcal{A}_5} \frac{|P(x_C) - \frac{N(x_C)}{|S|}|}{\frac{N(x_C)}{|S|}} = \frac{100}{|\mathcal{A}_5|}\sum_{x_C \in \mathcal{A}_5} \frac{|P(x_C)|S| - N(x_C)|}{N(x_C)}, \tag{20}$$

which is more sensitive in this respect since the same absolute difference of frequencies is less important if the empirical frequency $N(x_C)$ is high and more important for lower $N(x_C)$.

We have used the criteria E_a and E_r to evaluate the accuracy of the final distribution mixture (2). Table 3 contains the results obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and, for comparison, to the list $\mathcal{A}_4$ of properties specified by maximally four responses (second column). For both of the considered tests, Table 3 shows the mean relative and mean absolute error and the corresponding standard deviations. In addition we have computed the maximum relative and absolute errors and also the number of relative errors exceeding 100%. The mean relative error was 4.2% in the case of the list $\mathcal{A}_5$ and 4.1% in the case of the list $\mathcal{A}_4$; the corresponding absolute error was 338 and 460 respondents, respectively. Since all other results are also comparable, we may assume that a more extensive test would not yield substantially different values. We recall that by combining more than five responses we would mostly obtain very small frequencies $N(x_C)$ that would fall below the threshold N_ε, and therefore the resulting list would not be much longer than $\mathcal{A}_5$.

Let us recall that the relative error in the criterion E_r is invariant with respect to arbitrary norming. Consequently, the mean error of any displayed histogram column is 4.17%.

Table 3. Mean relative and mean absolute error of the statistical model with $M = 15{,}000$ components. Results obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and to the list $\mathcal{A}_4$ of properties specified by a maximum of four responses (second column)

List of test combinations:	$\mathcal{A}_4$	$\mathcal{A}_5$
Mean relative error in %:	4.07	4.17
Standard deviation of the relative error:	6.33	5.80
Maximum relative error of the model in %:	240.84	240.84
Number of relative errors exceeding 100%:	925	4,092
Mean absolute error:	470	348
Standard deviation of the absolute error:	951	655
Maximum absolute error of the tested:	45,779	45,779
Number of combinations tested:	3,468,134	26,425,727

In order to illustrate the distribution of relative errors in more detail we include Table 4. As can be seen, for very small empirical frequencies ($1{,}612 < N(x_C) < 3{,}000$) the mean relative error is 6.10% and quickly decreases for greater values of $N(x_C)$ (larger subpopulations). In the case of estimates based on the 10%-subset of microdata (last column) the first value 5.16% is smaller but the distribution of relative errors is similar. Our interactive software disallows evaluation of subpopulations $S(x_C)$ smaller than the threshold value $N_{0.05} = 1{,}612$ (see Appendix II) and indicates any histogram column that corresponds to a subthreshold frequency.

Obviously, the results in Table 3 strongly depend on the chosen subpopulation threshold N_ε. It is therefore unclear whether the achieved mean relative error 4.2% is to be considered too high or low enough. To answer this question we have compared the accuracy of our mixture model with the reproduction accuracy of a randomly chosen subset of 1 million individual microdata records (10% of S). Note that the standard method of disseminating statistical information by means of subsets of anonymized microdata provides the same comfort and flexibility as does the interactive statistical model. In particular, we can estimate the empirical frequencies $N(x_C)$ by using a representative subset of microdata. For the sake of comparison with the statistical model, we have evaluated the accuracy of the microdata subset in the same way as in Table 3. As we can see in Table 5, the accuracy of the 10% microdata subset is marginally better than the statistical model at reproducing the empirical frequencies. However, if the randomly chosen subset of microdata records should be used as a public-use file, it would be necessary to apply some sort of anonymization procedure to the data. Probably, after

Table 4. Distribution of relative errors of estimates according to the empirical frequency $N(x_C)$ (subpopulation size). Comparison of the statistical model and 10%-subset of microdata. In the first two columns we specify the lower and upper bounds of the frequency intervals, respectively. The third column contains the number of properties falling into the given interval of empirical frequencies. The last two columns contain the corresponding mean relative errors for the statistical model and subset of microdata, respectively

Interval	Lower bound	Upper bound	Number of combinations	Mean relative error in %	Mean relative error in %
1.	1,612	3,000	7,688,027	6.10	5.16
2.	3,000	5,000	5,011,625	4.88	3.86
3.	5,000	7,500	3,220,931	4.04	3.07
4.	7,500	10,000	1,906,156	3.50	2.58
5.	10,000	15,000	2,213,787	3.04	2.17
6.	15,000	30,000	2,695,817	2.38	1.67
7.	30,000	50,000	1,296,118	1.80	1.23
8.	50,000	100,000	1,075,615	1.37	0.94
9.	100,000	150,000	372,570	1.03	0.70
10.	150,000	300,000	358,112	0.78	0.55
11.	300,000	500,000	125,103	0.55	0.39
12.	500,000	1,000,000	71,104	0.39	0.28
13.	1,000,000	1,500,000	15,324	0.29	0.20
14.	1,500,000	3,000,000	8,511	0.22	0.14
15.	3,000,000	5,000,000	1,349	0.12	0.08
16.	5,000,000	10,300,000	200	0.02	0.04

Table 5. Relative and absolute accuracy of the estimates computed from the randomly chosen subset of 10% of microdata from the original database S. The test has been obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and to the list $\mathcal{A}_4$ (second column) in the same way as in the Table 3

List of test combinations:	$\mathcal{A}_4$	$\mathcal{A}_5$
Mean relative error in %:	2.94	3.60
Standard deviation of the relative error:	3.00	3.23
Maximum relative error of the model in %:	35.19	36.34
Number of relative errors exceeding 100%:	0	0
Mean absolute error:	307	409
Standard deviation of the absolute error:	450	1,913
Maximum absolute error of the model:	12,348	59,815
Number of combinations tested:	3,503,448	26,425,727
Number of microdata in the subset:	1,022,666	1,022,666

anonymization, the accuracy of the resulting public use file would decrease due to errors introduced by the anonymization process. In our comparison experiment the anonymization of microdata has been omitted.

5. Model-Based Information Analysis

Another possible way to utilize the latent information potential of the statistical model is to analyze the properties of subpopulations (cf. Grim et al. 2004a, b). A natural basis of information analysis is a virtual list $\mathcal{A}$ of statistically relevant subpopulations, which can be specified by combining variables (cf. (18)). The general scheme of the considered information analysis can be summarized as follows: we order, e.g., the virtual list $\mathcal{A}_4$ of 3.5 million statistically relevant subpopulations (specified by combinations of responses) according to a chosen statistical criterion and display the ordered list to the user. In some cases the ascending ordering of subpopulations (instead of the descending one) could also be of interest. In this section we suggest some criteria which may be useful for different purposes.

A very simple criterion applicable to ordering the subpopulations $\mathcal{A}$ is the conditional probability of a specific value $x_n \in \mathcal{X}_n$. We can order the subpopulations $S(x_C)$ from the list $\mathcal{A}$ according to the highest conditional probability $P_n|_C(x_n|x_C)$ (cf. (9)). By displaying the initial part of the ordered subpopulation list we can identify, e.g., social groups or subpopulations which are particularly hit by unemployment, if the variable x_n defines unemployed respondents. Obviously, we should exclude from evaluation the "trivial" subpopulations $S(x_C)$ for which $n \in C$ since in these cases the probability $P_{n|C}(x_n|x_C)$ is trivially only 1 or 0.

A simple modification of the conditional distribution $P_{n|C}(x_n|x_C)$ is to use the unconditional probability

$$P_{nC}(x_n, x_C) = P_{n|C}(x_n|x_C)P(x_C) = \sum_{m=1}^{M} w_m p_n(x_n|m)F_C(x_C|m) \tag{21}$$

The preceding criterion can be easily generalized to a pair of specified values $x_n \in \mathcal{X}_n, x_r \in \mathcal{X}_r$:

$$P_{nr|C}(x_n, x_r|x_C) = \sum_{m=1}^{M} w_m(x_C)p_n(x_n|m)p_r(x_r|m) \tag{22}$$

In this way the subpopulations can be ordered with respect to the highest relative frequency of a pair of values, for example we can identify subpopulations with high unemployment among young people. Analogously a natural alternative to this criterion is to use the unconditional probability

$$P_{nrC}(x_n, x_r, x_C) = P_{nr|C}(x_n, x_r|x_C)P(x_C)\sum_{m=1}^{M} w_m p_n(x_n|m)p_r(x_r|m)F_C(x_C|m) \tag{23}$$

which corresponds to the estimated frequency $|S|P_{nrC}(x_n, x_r, x_C)$ of the values x_n, x_r, x_C. Again, in the evaluation process we should exclude the combinations x_C for which $n,r \in C$, because the corresponding conditional probabilities $P_{nr|C}(x_n, x_r|x_C)$ are equal to 1 or 0.

In some cases we could be interested in subpopulations where the conditional distribution of a variable is concentrated on an arbitrary single value (or a small subset of values). For example, we could look in general for subpopulations having a typical (prevailing) type of occupation. In such a case, a suitable choice would be to use the minimum entropy criterion

$$H_{x_C}(\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} - P_{n|C}(x_n|x_C) \log P_{n|C}(x_n|x_C) \tag{24}$$

In other words, in the subpopulations characterized by a low entropy $H_{x_C}(\mathcal{X}_n)$, the answer to the nth question is almost unique. Note that it would be rather difficult to identify such subpopulations by other means, e.g., by calculating the relative frequencies.

The statistical model also provides a general possibility of identifying dependence between categorial variables. Recall that the standard tool to characterize a relationship between two real random variables is the correlation coefficient computed by means of the expected value of the normalized product of the involved variables. Unfortunately, in cases of discrete nominal variables like eye color, profession, marital status, etc., the product of two variables is not defined and there is no generally acceptable way to introduce a reasonable definition.

One possibility available for analyzing the statistical dependence between nominal (qualitative) random variables is to use the statistical information. If $X_n, X_r, n, r, \in \mathcal{N}$ are two discrete random variables, then their mutual statistical information can be expressed by means of the Shannon formula

$$I(X_n, X_r) = H(X_n) + H(X_r) - H(X_n, X_r) \tag{25}$$

where $H(X_n)$, $H(X_r)$, $H(X_n, X_r)$ are the respective Shannon entropies:

$$H(X_n) = \sum_{x_n \in X_n} - P_n(x_n) \log P_n(x_n), \quad P_n(x_n) = \sum_{m=1}^{M} w_m p_n(x_n|m), \quad n \in \mathcal{N}, \tag{26}$$

$$H(X_n, X_r) = \sum_{x_r \in X_r} \sum_{x_n \in X_n} - P_{nr}(x_n, x_r) \log P_{nr}(x_n, x_r), \quad n, r \in \mathcal{N}, \tag{27}$$

$$P_{nr}(x_n, x_r) = \sum_{m=1}^{M} w_m p_n(x_n|m) p_r(x_r|m). \tag{28}$$

The value of the Shannon information is zero if the two variables X_n, X_r are statistically independent and it is maximum if one of the two variables uniquely defines the value of the other one. The information criterion (25) can be used, e.g., to order the subpopulation list $\mathcal{A}$ according to the statistical dependence between two chosen variables.

6. Concluding Remarks

To date, one of the most informative known ways to disseminate statistical information is to release representative subsets of anonymized microdata. With appropriate microdata the users have the full freedom to examine arbitrary hypotheses and issues beyond the usual scope of data providers. Unfortunately, both the choice of a subset of the original microdata (typically about one million individual records) and the indispensable anonymization procedures may negatively influence the statistical validity of the contained information. Moreover, there is always some residual risk of disclosure and, for this reason, statistical agencies release microdata for research purposes only, usually under special license agreements and through secure data archives.

In view of these facts, the primary purpose of the considered statistical model is to make the census results freely available in a new, user-friendly way with a guaranteed confidentiality of data. However, the high level of confidentiality protection suggests also other application areas like medical registers and databases. The resulting interactive software provides flexibility and user comfort analogous to those sets of anonymized microdata at a comparable or even higher level of accuracy. In addition, the analytical simplicity of the underlying distribution mixture opens up new possibilities of information-oriented data analysis (data mining) based on efficient evaluation of a virtual list of several hundred thousands subpopulations.

Appendix I. Probabilistic Inference Mechanism

Note that any marginal distribution of the mixture (2) is easily obtained by deleting superfluous terms in the products $F(x|m)$. Actually, in view of this property, the discrete distribution mixture (2) can be used as a knowledge base of the Probabilistic Expert System (PES). Considering a basic situation, we assume $v_{i1}, v_{i2}, \ldots, v_{ik}$ to be a subset of input variables. Then, for any given input vector

$$x_C = (x_{i1}, x_{i2}, \ldots, x_{ik}) \notin \mathcal{X}_C, \quad C = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, N\}$$

and an output variable x_n, $(n \notin C)$, we can directly write equations for the related marginals

$$P_C(x_C) = \sum_{m=1}^{M} w_m F_C(x_C|m), \quad F_C(x_C|m) = \prod_{i \in C} p_i(x_i|m), \quad x_C \in \mathcal{X}_C, \tag{29}$$

$$P_{n,C}(x_n, x_C) = \sum_{m=1}^{M} w_m F_{n,C}(x_n, x_C|m), \quad F_{n,C}(x_n, x_C|m) = \prod_{i \in C \cup \{n\}} p_i(x_i|m), \tag{30}$$

and for the desired conditional distribution

$$P_{n|C}(x_n|x_C) = \frac{P_{n,C}(x_n, x_C)}{P_C(x_C)} = \sum_{m=1}^{M} W_m(x_C) p_n(x_n|m), \quad (P_C(x_C) > 0), \tag{31}$$

Here $W_m(x_C)$ are the component weights corresponding to the given input vector $x_C \in \mathcal{X}_C$:

$$W_m(x_C) = \frac{w_m F_C(x_C|m)}{\sum_{j=1}^{M} w_j F_C(x_C|j)}. \tag{32}$$

The conditional distributions $P_n|_C(x_n|x_C)$ (conditional histograms) describe the statistical properties of the subpopulation specified by the subvector x_C in terms of all variables x_n not included in x_C.

The interactive inference mechanism of the expert system PES can be extended to a more general case when each of the input variables x_{i_j} is confined to a subset of values

$$x_{i_j} \in \mathcal{D}_{i_j} \subset \mathcal{X}_{i_j}, \quad i_j \in C. \tag{33}$$

Thus, instead of an input subvector $x_C \in \mathcal{X}_C$, we are given a subset of input vectors $\mathcal{D}_C \subset \mathcal{X}_C$:

$$\mathcal{D}_C = \mathcal{D}_{i_1} \times \mathcal{D}_{i_2} \times \ldots \times \mathcal{D}_{i_k} \subset \mathcal{X}_C. \tag{34}$$

Analogously to Equations (29)–(30), we can write

$$F_C(\mathcal{D}_C|m) = \sum_{x_C \in \mathcal{D}_C} F_C(x_C|m) = \sum_{x_{i_1} \in \mathcal{D}_{i_1}} \cdots \sum_{x_{ik} \in \mathcal{D}_{i_k}} \prod_{i_j \in C} p_{i_j}(x_{i_j}|m) = \prod_{i_j \in C} p_{i_j}(\mathcal{D}_{i_j}|m), \tag{35}$$

and finally we can compute the conditional probability $P_{n|C}(x_n|\mathcal{D}_C)$ by

$$P_{n|C}(x_n|\mathcal{D}_C) = \frac{P_{n,C}(x_n, \mathcal{D}_C)}{P_C(\mathcal{D}_C)} = \sum_{m=1}^{M} W_m(\mathcal{D}_C) p_n(x_n|m). \tag{36}$$

Here $W_m(\mathcal{D}_C)$ are the component weights corresponding to the given subset of input vectors $\mathcal{D}_C \subset \mathcal{X}_C$:

$$W_m(\mathcal{D}_C) = \frac{w_m F_C(\mathcal{D}_C|m)}{\sum_{j=1}^{M} w_j F_C(\mathcal{D}_C|j)}. \tag{37}$$

The conditional distribution (31) represents an exact response to the definite input $v_C = x_C \in \mathcal{X}_C$. In a case of uncertain input information which is generally described by the probability distribution $\tilde{P}_C(x_C)$ on the subspace of input variables $\mathcal{X}_C$, we obtain

$$\tilde{P}_n(x_n) = \sum_{\mathbf{x}_C \in \mathbf{X}_C} P_{n|C}(x_n|\mathbf{x}_C)\tilde{P}_C(\mathbf{x}_C) = \sum_{m=1}^{M} \tilde{W}_m p_n(x_n|m), \tag{38}$$

where

$$\tilde{W}_m = \sum_{\mathbf{x}_C \in \mathbf{X}_C} W_m(\mathbf{x}_C)\tilde{P}_C(\mathbf{x}_C). \tag{39}$$

Let us remark that Formula (38) realizes the so-called memoryless information channel with noise – a well-known object of information theory.

The processing of uncertain information is a typical feature of expert systems as decision-supporting tools. In the present context, the uncertain information on input can be used to analyse the properties of hypothetical subsets or subpopulations. Note that the choice of input and output variables is affected only by their meaning or availability – without any formal restrictions implied by the knowledge base. Thus, unlike the rule-based systems, the expert system PES is fully symmetrical with respect to the role of variables.

Appendix II. Statistical Validity of Census Data

The standard census is a statistical investigation of extreme extent that includes the entire population. All individuals have to answer a set of questions and for any combination of possible answers we can specify the exact number of respondents having the corresponding property. This number is fixed and uniquely given by the current state of the population. Nevertheless, for theoretical reasons, any statistical property of the census population specified by means of empirical frequency has a limited validity.

Obviously, a census is unique and cannot be repeated as a random experiment under identical conditions. On the other hand, the census questionnaire can be viewed formally as a vector of discrete finite-valued random variables

$$v = (v_1, v_2, \ldots, v_N), \quad v_n \in \mathcal{X}_n,$$

and every respondent can be assumed to provide an independent observation x of this random vector

$$x = (x_1, x_2, \ldots, x_N) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N.$$

From a theoretical point of view, statistical properties of a random vector can be described in full generality by a joint probability distribution $P*(x)$ of the involved random variables. We can assume that for a given property specified by a combination of values $x_C = (x_{i1}, \ldots, x_{ik})$ there is an unknown "hypothetical" probability

$$p* = P*(x_C), \quad x_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \ldots \times \mathcal{X}_{i_k} = \mathcal{X}_C \tag{40}$$

of this property. The probability $P*(x_C)$ represents a statistical property of the subpopulation characterized by x_C and can be estimated by means of the corresponding relative frequency. Typically, speaking about the statistical properties, we refer to large subpopulations; let us recall that a statistical property becomes questionable if the related subpopulation is small.

We can formally view the number of respondents s_n satisfying the considered property x_C as a random result of n independent observations of a random event (so-called Bernoulli trials) with probability $p* = P*(x_C)$. We recall that, in the case of a census, the number s_n is fixed and uniquely defined and $P*(x_C)$ is a hypothetical and unknown probability which cannot be verified by repeated experiments. Despite these facts, the relationship between the observed number of respondents s_n with the property x_C, the population size n and the probability $P*(x_C)$ can be used to characterize the reliability of the empirical frequency s_n. In other words, we can decide whether the statistical property is statistically significant or not. We recall first the well-known De Moivre-Laplace limit theorem of probability in the following simple form (cf. Feller 1962):

Theorem. Let s_{n} stand for the number of successes in n Bernoulli trials with probability $p*$, $(0 < p* < 1)$ of success. Then for each fixed α, $(0 < \alpha < 1)$ the probability

$$P\left\{|\frac{s_n}{n} - p*| < \alpha\right\} = P\left\{|\frac{s_n - np*}{\sqrt{np*(1-p*)}}| < \alpha\sqrt{\frac{n}{p*(1-p*)}}\right\} \tag{41}$$

satisfies the relation

$$\lim_{n\to\infty}\left[P\left\{|\frac{s_n}{n} - p*| < \alpha\right\} - \frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p*(1-p*)}}} \exp\left(-\frac{z^2}{2}\right)dz\right] = 0. \tag{42}$$

This theorem provides an asymptotic approximation of the probability that the observed relative frequency s_n/n of a random event and the related unknown probability $p*$ differ from each other by less than a fixed upper bound α, $(0 < \alpha < 1)$. From the theorem it follows that for large values of n the error bound probability (41) can be approximated by the integral expression:

$$P\left\{|\frac{s_n}{n} - p*| < \alpha\right\} \approx \frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p*(1-p*)}}} \exp\left(-\frac{z^2}{2}\right)dz. \tag{43}$$

In particular, if we intend to clarify the conditions implying that Probability (41) is high enough, e.g., at the confidence level 0.95, we have to guarantee the inequality

$$P\left\{|\frac{s_n}{n} - p*| < \alpha\right\} \geq 0.95, \quad (0 < \alpha < 1). \tag{44}$$

In other words, from Inequality (44) it follows that (at the confidence level 0.95) the unknown true probability $p*$ belongs to the interval

$$\frac{s_n}{n} - \alpha < p* < \frac{s_n}{n} + \alpha \tag{45}$$

or, equivalently, the relative frequency s_n/n falls between the bounds

$$p* - \alpha < \frac{s_n}{n} < p* + \alpha. \tag{46}$$

In view of the limit theorem we can analyze, for large values of n, the inequality

$$\frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p*(1-p*)}}} \exp\left(-\frac{z^2}{2}\right)dz \geq 0.95. \tag{47}$$

The integral on the left-hand side (sometimes called the error function) is an increasing function of the upper bound. It can be proven that inequality (47) is satisfied if it holds that

$$\alpha\sqrt{\frac{n}{p*(1-p*)}} \geq 1.96. \tag{48}$$

Now, as we are interested in the relative accuracy of estimates, we choose the threshold α in the dependence of $p*$, e.g., by setting $\alpha = ap*$, $(0 < a < 1)$. Here $a = \alpha/p*$ is the required relative accuracy of the probability $p*$. Substituting for α in (46) we can write

$$\frac{|\frac{s_n}{n} - p*|}{p*} = \frac{|s_n - np*|}{np*} < a \tag{49}$$

or equivalently

$$(1-a)np* < s_n < (1+a)np*. \tag{50}$$

Making substitution for α in (48), we obtain

$$a\sqrt{\frac{np*}{(1-p*)}} \geq 1.96 \tag{51}$$

and further we obtain the lower bound for the probability $p*$ as a function of its required accuracy a:

$$p* \geq \frac{(1.96)^2}{(1.96)^2 + a^2 n}. \tag{52}$$

In other words, if we want to guarantee the relative accuracy a, $(0 < a < 1)$, then the estimated probability $p*$ must be greater than the expression on the right hand side. In particular, for the Czech Census 2001 we can fix the number of respondents to $n = 10{,}230{,}060$ and, by substitution, we can describe the underlying relation between $p*$ and a as in Table 6.

If we choose the "admissible" relative error $a = 0.05$ then, for the true unknown frequency satisfying the inequality

$$np* > 1{,}536 \tag{53}$$

Table 6. Relationship between the relative accuracy a of the probalility p of a specific property of respondents, and of the related subpopulaton size np* (at the confidence level 0.95), in the case of a census population of size 10,230,060*

a	0.01	0.02	0.05	0.10
$p* >$	0.0037412	0.0009379	0.0001502	0.0000376
$np* >$	38,272	9,595	1,536	384

the corresponding empirical frequency s_n is bounded by the inequality (cf. (49))

$$0.95np* < s_n < 1.05np*. \tag{54}$$

From inequality (54) it follows that, in turn, the unknown frequency $np*$ is bounded by means of the empirical frequency s_n:

$$\frac{s_n}{1.05} < np* < \frac{s_n}{0.95}. \tag{55}$$

However, the last inequality holds only if $np* > 1{,}536$. As the true unknown frequencies $np*$ are not available, we consider only the empirical frequencies s_n which guarantee the above condition (53). In particular, we confine ourselves to the empirical frequencies s_n satisfying the inequality

$$\frac{s_n}{1.05} > 1{,}536 \tag{56}$$

which implies the condition (53)

$$np* > \frac{s_n}{1.05} > 1{,}536 \tag{57}$$

Consequently, if the empirical frequency s_n is greater than 1,612:

$$s_n > 1.05 * 1{,}536 = 1{,}612.8, \tag{58}$$

then the unknown true frequency $np*$ is greater than 1,536 (cf. (53)) and cannot differ from s_n by more than 5%. We can conclude that, by using the lower bound $N_{0.05} = 1{,}612$, we guarantee the 5% accuracy of the empirical frequencies s_n (at the confidence level 0.95) for the census population size $|S| = 10{,}230{,}060$.

7. References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. Statistisk Tidskrift, 15, 429–444.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Duncan, G. and Lambert, D. (1989). The Risk of Disclosure for Microdata. Journal of Business & Economic Statistics, 7, 207–217.

Feller, W. (1962). An Introduction to Probability Theory and Its Applications, I. New York, London: John Wiley & Sons.

Fienberg, S.E. (1994). Conflicts between the Needs for Access to Statistical Information and Demands for Confidentiality. Journal of Official Statistics, 10, 115–132.

Fienberg, S.E., Makov, U.E., and Steel, R.J. (1998). Disclosure Limitation using Perturbation and Related Methods for Categorical Data, with Discussion. Journal of Official Statistics, 14, 485–502.

Grim, J. (1982). On Numerical Evaluation of Maximum – Likelihood Estimates for Finite Mixtures of Distributions. Kybernetika, 18, 173–190.

Grim, J. (1990). Probabilistic Expert Systems and Distribution Mixtures. Computers and Artificial Intelligence, 9, 241–256.

Grim, J. (1992). A Dialog Presentation of Census Results by Means of the Probabilistic Expert System PES. In Proceedings of the Eleventh European Meeting on Cybernetics and Systems Research, R. Trappl (ed.). Singapore: World Scientific, Vienna 21–24 April, 997–1005.

Grim, J. (1994). Knowledge Representation and Uncertainty Processing in the Probabilistic Expert System PES. International Journal of General Systems, 22, 103–111.

Grim, J. and Boček, P. (1996). Statistical Model of Prague Households for Interactive Presentation of Census Data. In SoftStat'95. Advances in Statistical Software 5. Stuttgart: Lucius & Lucius, 271–278.

Grim, J., Boček, P., and Pudil, P. (2001). Safe Dissemination of Census Results by Means of Interactive Probabilistic Models. In Proceedings of the ETK-NTTS 2001 Conference, P. Nanopoulos and D. Wilkinson (eds). Rome: European Communities, 849–856.

Grim, J., Hora, J., Boček, P., Somol, P., and Pudil, P. (2004a). Information Analysis of Census Data by Using Statistical Models. In Proceedings: Statistics – Investment in the Future. Prague.

Grim, J., Hora, J., and Pudil, P. (2004b). Statistical Model for Interactive Presentation of Census Results under Protection of Confidentiality. Statistika, 40(5), 400–414, [In Czech]

McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models. New York, Toronto: John Wiley & Sons.

Schlesinger, M.I. (1968). Relation between Learning and Self-learning in Pattern Recognition. Kibernetika, (Kiev), (2), 81–88, [In Russian]

Willenborg, L.C.R.J. and de Waal, A.G. (2001). Elements of Statistical Disclosure Control. New York: Springer Verlag.

Williams, T. (2005). The Development of Truth Decks for the 2010 Census Count Imputation Research. In Proceedings of the American Statistical Association, Section on Survey Research Methods.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Research in Official Statistics, 2, 87–104.

Received April 2009
Revised December 2009