MIXTURE MODELS OF MISSING DATA

Tamás Rudas

Eötvös Loránd University and

TÁRKI Social Research Institute

rudas@tarki.hu

Department of Statistics

Faculty of Social Sciences

Eötvös Loránd University

Pázmány Péter sétány 1/A

H-1117 Budapest, Hungary

Fax: + 36 1 372 2912

MIXTURE MODELS OF MISSING DATA

This paper proposes a general framework for the analysis of survey data with missing observations. The approach presented here treats missing data as an unavoidable feature of any survey of the human population and aims at incorporating the unobserved part of the data into the analysis rather than trying to avoid it or make up for it. To handle coverage error and unit nonresponse, the true distribution is modeled as a mixture of an observable and of an unobservable component. Generally, for the unobserved component, its relative size (the no-observation rate) and its distribution are not known. It is assumed that the goal of the analysis is to assess the fit of a statistical model, and for this purpose the mixture index of fit is used. The mixture index of fit does not postulate tat the statistical model of interest is able to account for the entire population rather, that it may only describe a fraction of it. This leads to another mixture representation of the true distribution, with one component from the statistical model of interest and another unrestricted one. Inference with respect to the fit of the model, with missing data taken into account, is obtained by equating these two mixtures and asking, for different no-observation rates, what is the largest fraction of the population where the statistical model may hold. A statistical model is deemed relevant for the population, if it may account for a large enough fraction of the population, assuming the true (if known) or a sufficiently small or a realistic no-observation rate.

Key words: missing data, mixture index of fit, model diagnostics, no-fit rate, no-observation rate

1. INTRODUCTION

Most sample surveys yield data sets that are incomplete in various ways and for various reasons. The current attitude and methodology to handle the missing information issue can be summarized as trying to avoid it in the data collection phase, trying to make up for it during initial data screening, and pretending it never existed, in the final analysis. Some people will always refuse to participate in the survey, some, although selected into the sample, will not be found and many will be missed by the sampling procedure, giving them unintentionally too little or no chance at all of being selected. Some others, although they are available for the survey, choose not to respond to certain questions, leading to a different kind of missing information. All that is far too important and potentially informative, to be considered a nuisance.

This paper argues that the correct attitude to missing information is to model it, and one such approach is the following. It is assumed that for every survey the entire population may be divided into two groups, one consisting of those who are available to participate in the survey and another one, consisting of those who are not. The distribution of the relevant characteristic among those who are available for the survey can be observed but the distribution in the other, non-available, part cannot be observed. Therefore, the distribution in the entire population is a mixture of two distributions, say, O in the observable part and U in the unobservable part:

$$(1-\rho)O + \rho U,$$

where 1-ρ is the relative size of the fraction of the population that is available for observation and ρ is the relative size of the fraction not available for observation, for some ρ between 0 and 1.

Very often, the goal of the survey is to test a hypothesis that assumes that the true distribution of the population belongs to a subset *H* of all possible distributions, that is, a statistical model defined by the presence of some relevant property. This hypothesis may only be true for a part of the population, where the distribution of the characteristic of interest belongs to *H,* but not necessarily for the entire population. Let (1-π) denote the relative size of the part where the hypothesis is true and *F* the distribution there. The remaining part of the population is described by another distribution, say *E.* Then, the true distribution is a mixture

$$(1-\pi)F + \pi E,$$

for some π between 0 and 1 and the larger is the fraction where *F* is the distribution, the better is the fit of the model. This is the idea underlying the mixture index of fit (Rudas, Clogg, Lindsay, 1994, Clogg, Rudas, Xi, 1995).

Because both of the mixtures above describe aspects of the same distribution,

$$(1-\rho)O + \rho U = (1-\pi)F + \pi E,$$

and the smaller are ρ and π, the better is the fit of model *H*: if ρ is small, the analysis is based on data from a large fraction (1- ρ) of the population and if π is small, the model can be assumed to describe a large fraction (1- π) of the population. The problem of minimizing ρ and π was considered by Xi (1996), but in a different context. One attractive feature of this approach is that it is not restrictive, that is, it always holds true for some values of ρ and π. Therefore, the mixture representation is a framework, not a model, for the analysis of data with missing information. The traditional approach to fitting statistical models to data is essentially solving the equation above in F approximately, under the assumption that both ρ and π are zero, that is, under the assumption that the observations are complete (and were selected without restrictions from the entire population) and that the model describes the entire population. Whenever any of these assumptions seems unrealistic, the approach of the present paper can be applied.

The method proposed for the analysis of data sets with missing information is described in section 2 and is based on the estimate, for given values of ρ, of the smallest value of π, say $\pi(\rho)$, for which a representation of the form above is possible. The selected values of ρ may be derived from the known, estimated or acceptable values of the no-observation rate and $\pi(\rho)$ is the estimate of the smallest possible fraction of the population outside of the model assuming the given no-observation rate. In addition to describing the possible inferential procedures in the framework, useful properties of the function $\pi(\rho)$ are discussed in section 3.

The analysis based on the mixture representation is appropriate to handle coverage error, no-contact and unit nonresponse. The framework may also be extended to include item nonresponse as well by including further components into the mixture representation of the observational procedure to model parts of the population that were available to respond to some but not all of the questions. This extension, however, is not discussed in the paper. Finally, section 4 summarizes the results presented in the paper and points to problems that require further research.

The inferential procedures within the present framework are illustrated by new analyses of the Blau, Duncan (1967) mobility data and data from the International Social Survey Programme 1995 module on national identity (ISSP, 1995). For the father-son mobility data, the goodness of fit of the models of independence, quasi-independence and quasi-uniform association is discussed. The results show, that even if small to moderate no-observation rates are assumed, the fits of these models appear considerably better than thought before. For the ISSP data, the independence between proudness of results in sport and of the way democracy works is investigated using the $\pi(\rho)$ function and the effect of item nonresponse is also analyzed.

2. THE MIXTURE APPROACH TO HANDLING MISSING DATA

There are various factors influencing the ability and willingness of every person in the population of interest to contribute information to a survey. The effects of these factors cannot always be separated but some parts of the missing information are not collected because of errors and some other parts of missing information are not collected because of inherent characteristics of the population (like no chance of being contacted or lack of willingness to participate, if contacted). The effects of these factors are best modeled by assuming that every person in the population of interest has a certain probability of participating in the survey. This probability may depend on the selection probability associated with the sampling procedure, the conditional probability that a person will be contacted, given that he or she was selected into the sample and the conditional probability that the person is ready to participate if contacted. All these probabilities may depend on the methodology of the survey, on the way in which the procedures are carried out, on the topic of the survey and possibly on other factors as well and may be different for every person in the population.

This approach, although conceptually complete, is too complicated for any practical data analytic application and therefore, the framework proposed in the present paper is a simplification of it. It is assumed that for every survey, some people are available to participate, some are not. Those who are given zero (or practically zero) selection probabilities in the sampling procedure, those who (usually) cannot be found after the given number of attempts to contact, those who would choose not to participate in he survey even if contacted successfully, are not available for the survey. The other members of the population are available for the survey. In other words, the simplification considered here assumes that the probability of availability is 0 for some and 1 for the others. These probabilities, of course, also depend on the actual survey. A person who is not available to participate in a survey, may be available for another one, which is done by different methods, on a different topic, by another survey organization. The approach proposed here does not make assumptions with respect to the reasons that make a person not available for a survey. Also, no distinction is made between non-availability due to errors in the design or its implementation, and non-availability due to inherent characteristics of a person, because these causes cannot always be separated and are irrelevant from the point of view of inference based on the actually observed data. Further, no attempt will be made to identify the fraction of missing information due to methodological errors. Rather, statistical models will be fitted with the missing information taken into account.

More formally, let ρ be the relative size of the fraction of the population that is not available for participation and $1-\rho$ the relative size of the fraction that is available for participation. The value of ρ is unspecified for the time being, representing the fact that although one can assume the existence of these two parts in the population, even with a well specified data collection methodology, one has no information about their respective sizes before the data are collected. The value of ρ lies between 0 and 1. The value 0 describes the situation when everybody is available to participate in the survey and the value 1 describes the situation when nobody is available to participate. Realistic values of ρ lie in between these extreme values. Note that the unit nonresponse rate cannot be used to determine the value of ρ, because our concept of not being available for participation includes coverage errors and no-contacts as well, and the size of coverage error is usually difficult or impossible to estimate. This however, does not mean that ρ is greater than the observed nonresponse rate: ρ is the no-observation rate, that is, the overall probability that someone cannot be observed, while the nonresponse rate is the conditional probability of nonresponse among those who were contacted and there is no generally valid inequality between these two quantities.

It may be assumed without restriction that the primary goal of the survey is to estimate the joint distribution of certain variables in the population and a further goal may be to test prespecified hypotheses with respect to the distribution of these variables. The distribution of interest can, of course, only be observed and estimated on the part of the population that is available for participation. Let O denote this observable distribution. One has to assume, that in the other part, that cannot be observed, some other distribution, say, U is valid. An analysis disregarding missing information would assume that the two distributions O and U are the same. Such an assumption will not be made here rather, U will be allowed to be different from O. The foregoing can be represented by saying that the distribution in the population is a mixture of the two distributions O and U and the mixing weights are the relative sizes of the fractions where these distributions hold true.

$$(1-\rho)\, O + \rho\, U. \tag{1}$$

It should be noted that (1) does not restrict reality: whatever is the size of the fraction not available for observation (including the two extreme situations), there will be a value of ρ, with which (1) holds true. Therefore, (1) is not a model for missing information rather, it is a framework in which missing information can be handled (as opposed to the usual approach that neglects its existence).

The framework summarized in (1) does not make it possible to estimate ρ and the distribution *U*. This will be done in parallel with considering a hypothesis *H* which the researcher wishes to test based on the available data. Such a hypothesis assumes that the true distribution belongs to a model, that is a subset (usually defined by the presence of some desirable property) of all possible distributions.

With respect to the fit of the model, the population can also be divided into two parts: in one of them, the hypothesis is true and it may not be true in the other one. This is the idea underlying the mixture index of fit proposed by Rudas, Clogg, Lindsay (1994). Let (1-π) denote the relative size of the fraction where the model fits and π that of the fraction where the model does not fit. This leads to the following mixture representation of the true distribution:

$$(1-\pi)\, F + \pi\, E, \tag{2}$$

where *F* is a distribution in the model and *E* is another, not specified distribution. The larger is the size of the fraction 1-π, the better is the fit of the model. The smallest possible value of π which makes the representation (2) possible is the value of the mixture index of fit. For justification, properties and applications of the mixture index of fit see also Clogg, Rudas, Xi (1995), Clogg, Rudas, Matthews (1998) and Rudas, Zwick (1997). Algorithmic aspects were discussed by Xi, Lindsay (1996) and generalizations by Formann (2000, 2003), Knott (2001) and Rudas (1998, 1999). Rudas (2002) gives a non-technical overview of the use and interpretation of the mixture index of fit and many of the properties discussed in those papers apply to the methodology presented here. The most important properties of the mixture index of fit include that it is defined in a non-restrictive framework, its estimated value does not depend on the sample size in the way chi-squared based indices do and it has a straightforward interpretation. The mixture index of fit is the smallest fraction of the population which is inconsistent with hypothesis *H* and methods for point and interval estimation of its value are available.

Both equations (1) and (2) represent the true distribution in the entire population, one from the aspect of missing information, the other one from the aspect of the ability of a particular statistical model to account for it. As the final goal of the survey is to test *H*, the two representations of the true distribution are equated with each other to see how appropriate is hypothesis *H* to account for the population, in light of the survey data if the existence of missing information is also taken into account. This leads to the following equation which serves as the basis of assessing model fit in the presence of missing information.

$$(1-\rho)\, O + \rho\, U = (1-\pi)\, F + \pi\, E. \qquad (3)$$

In (3), ρ is the fraction that was not observed and π is the fraction where the model does not fit. In an ideal situation, both these fraction are zero, that is, the entire population was (or could be) observed, and the hypothesis of interest holds in the entire population. In reality, both these fraction are different from zero. The smaller is ρ, the larger is the observed fraction on which the analysis is based and the more certain the result are. The smaller is π, the larger is the fraction of the population where the hypothesis is true. Therefore, whether the population of interest can be assumed to possess the property formulated in *H*, based on the present data, depends on whether both ρ and π are small. More precisely, the smaller are both these quantities, the less evidence is provided by the actual data against assuming that the property formulated in *H* is true for the population. For a different interpretation of (3) with respect to model fit see Xi (1996).

To illustrate the above framework, suppose the joint distribution of two binary variables is of interest and the observed distribution (that, for simplicity, is assumed to be a good estimate of the true distribution in the observable part of the population) is as given in Table 1.

***insert Table 1 around here ***

If the model of interest is independence of the two variables, (3) may take the form of the representation in Table 2. The representation in Table 2 uses four distributions. The first one is the observed distribution and this is the estimate of the true distribution in the observable part of the population. The second distribution is an assumption regarding the distribution in the part of the population that was not observed. This distribution says about the non-observable part that category (2,2) is more likely among those who were not available for the survey than among those who were, everybody in cell (1,2) was available to participate in the survey and among those who were not observed, cells (1,1) and (2,1) are equally likely. The third distribution is independent, that is, belongs to the statistical model and the fourth distribution describes those, for whom the hypothesis of independence does not hold. This distribution says that failure of the independence hypothesis to describe the entire population is due to an excess number of people being in the second category of the first variable, and mostly to those who are also in the second category of the second variable.

*** insert Table 2 around here ***

The four constants in Table 2 have the following meanings: 80% of the population was available for observation and 20% was not; independence is able to account for 93% of the population and 7% of it lies outside of independence. The conclusion from Table 2 may be that having observed 80% of the population, we find a 93% fit of independence. This conclusion will be further discussed and justified later on. Either side of the representation in Table 2 yields the same distribution, presented in Table 3, for the entire population. This is obtained by carrying out the operations given in Table 2 in every cell of the 2x2 table.

*** insert Table 3 around here ***

The representation given in Table 2 is only one of the several possibilities. Having observed the distribution in Table 1, one may consider the representation given in Table 4. Here, again, the first table on the left hand side contains the observed distribution and on the right hand side an independent distribution. But the representation in Table 4 assumes that only 10% of the population was not available for observation and that 99% of the population can be described by independence. Therefore, Table 4 suggests a better fit of the independence model, because it may account for a larger fraction of the population, based on data from a larger fraction, than it was suggested by the representation in Table 2. Note, that the estimate for the distribution in the entire population that can be derived from Table 4, is different from the one given in Table 3, that was derived from Table 2 however, both representations in Tables 2 and 4 are compatible with the observed data.

*** insert Table 4 around here ***

What kind of conclusions are justified based on a representation of the form of (3), like the ones in Tables 2 and 4? The analyst may say that having observed a certain fraction of the population, the hypothesis of interest may account for a given fraction of the entire population. An analyst may take the position of not relying on data collected from less than a specified fraction of the population; or that a hypothesis is not relevant for a population unless it can describe at least a certain fraction of it. It may very well be the case that based on the representation in Table 2 someone concludes that as only 80% of the population was observed, no substantial conclusion can be drawn with respect to the hypothesis of independence. Based on Table 4, the conclusion may be that based on the observation of 90% of the population, it appears that only 1% of it cannot be described by independence.

As the foregoing example illustrated, in general, there are various values of ρ and π that are compatible with the observed distribution. One useful feature of the relationship between the possible (ρ, π) pairs, which follows from a basic property of the mixture index of fit, is that if a representation of the form (3) with (ρ, π) is possible, then a representation with any (ρ, π') is also possible, if $\pi \leq \pi' \leq 1$. That is, for the same no-observation rate, any value of the no-fit rate is possible that is greater than a possible no-fit rate. Therefore, for an observed distribution and for every no-observation rate, there is a smallest no-fit rate, with which a representation of the form (3) is possible. This value, for the given value of ρ will be denoted by $\pi(\rho)$. The value of $\pi(\rho)$ is obtained as the smallest value of the mixture index of fit for the hypothesis of interest and the distributions on the left hand side of (3), where the minimum is taken over all possible unobserved distributions *U,* for the given no-observation rate ρ.

To illustrate the inferential procedures facilitated by $\pi(\rho)$, the Blau, Duncan (1967) 5x5 father-son mobility table, as condensed in Knoke, Burke (1980), will be revisited here. Clogg, Rudas, Xi (1995) assessed the fit of the independence, quasi-independence (i.e., independence except for the main diagonal) and quasi-uniform association (i.e., constant local odds ratios, except for the main diagonal) models to these data using the mixture index of fit. The application of the mixture index of fit was especially appropriate to these data, as the real sample size, that has a great influence on chi-squared based analyses, is not known: the reported figures are population estimates in tens of thousands. The approach presented here to incorporate missing information is similarly insensitive to sample size. Clogg, Rudas, Matthews (1998) applied simple graphical techniques to visualize the results and to analyze the residuals. Similar techniques could be applied to the present, more general, approach as well.

*** insert Table 5 around here ***

Table 5 presents the values of the traditional Pearson chi-squared and likelihood ratio statistics, the mixture index of fit π^* and the values of $\pi(\rho)$ for different ρ values. The values of $\pi(\rho)$ and the estimates of the distributions *U, F,* and *E* were calculated using programs written in MATLAB (2002). The programs used to calculate the $\pi(\rho)$ values reported in Table 5 can be downloaded from Verdes (2002). These can be modified to perform similar calculations for other models and other data sets.

The mixture index of fit π^* indicates that one estimates at least 31% of the population to be outside of independence, quasi-independence may account for nearly 85% of the population and quasi-uniform association may hold for nearly 95% of the population. The analysis by assuming missing observations modifies this picture. If one assumes as little as 5% no-observation, independence is estimated to be able to describe nearly 80% of the population. By assuming a higher no-observation rate, the model appears to be able to account for larger fractions of the population. If a no-observation rate of 15% is assumed, independence is estimated to be valid on nearly 87% of the population. If the fact that information is likely to be missing from the data is taken into account, he data appear to provide the researcher with less evidence against the model of independence than if the researcher believes the data set to be complete. Similar interpretations can be given to the values of $\pi(\rho)$ for the other models. In fact, if one assumes missing information, all models appear to show a better fit than without thinking of the possibility of missing observations. Considering the quasi-uniform association model to describe these data is hardly justified, as the model of quasi independence shows a fairly good ability to describe a large part of the population, even with these modest no-observation rates. For example, if one assumes a 10% no-observation rate, quasi-independence may describe more than 94% of the population. Representation (3) for the case of 10% no-observation rate and the model of quasi-independence is given in Table 6. The categories, in the rows and columns, are professional and managerial, clerical and sales, craftsmen, operatives and laborers, farmers.

*** insert Table 6 around here ***

In the representation shown in Table 6, the observed distribution has a weight of 0.9 (the observation rate) and the not observed distribution has a weight of 0.1 (the no-observation rate). On the other side of the equation, the quasi-independent distribution has a weight of 0.944 (the fit rate) and the unrestricted distribution has a weight of 0.056 (the no-fit rate). In the representation given in Table 6, missing observations are mostly attributed to downward mobile sons in the third and fourth categories. In fact, nearly 90% of the missing observations belong to these categories. On the other hand, the lack of (complete) fit of the quasi independence model is mostly attributed to upward mobile sons in the first category, representing more than 70% of the total misfit. Sons in the first two categories account for more than 90% of the total no fit. The analysis implies, that by assuming a 10% no-observation rate, the model of quasi-independence may be valid for as much as 94.4% of the population and the misfit of the model can be attributed to “too few” observations in the third and fourth categories and “too many” observations in the first and second categories. Moreover, compared to a population which is completely described by quasi-independence, the data set contains “too few” downward mobile respondents and “too many” upward mobile respondents.

The foregoing analysis illustrates, that the representation cannot only be used to assess model fit, with the assumption of missing observations, but is also applicable to describe the actual data, in comparison to a statistical model. In the present example, one can conclude that the mobility process represented by the data contained more upward (and, consequently, less downward) mobility than one would expect under quasi-perfect mobility.

To achieve a final decision as to how relevant the model of quasi-independence may be to describe the population underlying the data, the analyst has to decide whether the no-observation rate and the estimated distribution in the part that was not available for observation are realistic and whether or not the estimated fit rate is sufficiently large. This is a much less formalized and much less automatic procedure than simply applying standard tests of fit and making decisions based on the achieved significance levels (the p-values). But this different approach to judging model fit seems appropriate in an environment, where the available data describe a fraction of the population only and the correct magnitude of and reasons for missing information are not precisely known.

3. THE $\pi(\rho)$ FUNCTION

In this section, $\pi(\rho)$ will be considered as a function of ρ, and properties of this function, which may be useful in assessing model fit, will be studied. Formally, $\pi(\rho)$ was defined as

$\pi(\rho) = \min(\pi$: there exist U, E unrestricted and F in H, such that $(1-\rho)O + \rho U = (1-\pi)F + \pi E)$.

The first property of $\pi(\rho)$ is that if the no-observation rate is 0, that is, the observations are supposed to have come from the entire population, then it is equal to the mixture index of fit:

$$\pi(0) = \pi^*.$$

Next, it will be shown that $\pi(\rho)$ is a monotone decreasing function, that is,

$$\pi(\rho') \geq \pi(\rho''), \text{ if } \rho' \leq \rho''.$$

To see this, notice that if there is a representation of the form

$$(1-\rho')O + \rho' U' = (1-\pi')F' + \pi' E', \tag{4}$$

then, for every $\rho'' \geq \rho'$, there is also a representation of the same form:

$$(1-\rho'')O + \rho'' U'' = (1-\pi')F' + \pi' E',$$

that is one, where ρ'' is used with π'. Indeed, U'' can be selected as

$$U'' = ((\rho'' - \rho')O + \rho' U') / \rho''.$$

It is easy to see that this U'' is a distribution, indeed. Therefore, the set of possible π values for ρ'' contains the set of possible π for ρ' and the minimum for ρ'' is smaller than or equal to the minimum for ρ', proving that $\pi(\rho)$ is a monotone decreasing function.

The former property implies that if $\pi(\rho') = 0$ then $\pi(\rho'') = 0$ if $\rho' \leq \rho''$, because the values of the function $\pi(\rho')$ are between 0 and 1.

For ρ values, such that $\pi(\rho)$ is positive, the function is also strictly monotone decreasing, that is,

$$\pi(\rho') > \pi(\rho''), \text{ if } \rho' < \rho'' \text{ and } \pi(\rho'') > 0 \,.$$

Notice, that in the condition above, $\pi(\rho'') > 0$ implies that $\pi(\rho') > 0$ and as ρ'' cannot exceed 1, $\rho'<1$. To see that this property holds, consider a representation as in (4). From this, another one will be derived, such that, on the left hand side, the weight of O is $(1-\rho'')$, and, on the right hand side, F' appears with a weight greater than $(1-\pi')$. Applying this to $\pi' = \pi(\rho')$ proves that among the possible π values for ρ'' there is one that is smaller than the smallest possible value of π for ρ' and therefore, $\pi(\rho'') < \pi(\rho')$. To obtain a representation with this property, let x be a (small) positive constant and add xF' to both sides of (4). Then, the sum (over the cells of the table) of either side of the equation will be 1+x. Now, if one normalizes by dividing both sides by 1+x, the weight of O becomes smaller than what it was and the weight of F' becomes greater than what it was. To obtain the desired weight of O, that is, $(1-\rho'')$, x should be selected as $x=(\rho''-\rho')/(1-\rho'')$ and by the conditions this is a strictly positive quantity. Then, from (4) it follows that

$$(1-\rho')O + \rho'U' + (\rho''-\rho')/(1-\rho'')\, F' = (1-\pi')F' + (\rho''-\rho')/(1-\rho'')\, F' + \pi'E'$$

or that

$$(1-\rho')O + \rho'U' + (\rho''-\rho')/(1-\rho'')F' = (1-\pi'+(\rho''-\rho')/(1-\rho''))F' + \pi'E'.$$

The normalization requires dividing both sides by $1+(\rho''-\rho')/(1-\rho'')=(1-\rho')/(1-\rho'')$. On the left hand side this leads to $(1-\rho'')O$ and another component that, because the sum is normalized to be 1, can be written as ρ'' times a distribution. On the right hand side, the coefficient of F' turns out to be $(1-\pi' + \rho''\pi')/(1-\rho')$. This is easily seen to be greater than $(1-\pi')$ and therefore, may be written as $(1-\pi'')$for some $\pi''<\pi'$, completing the proof.

An important consequence of the last property is that for those values of ρ, where $\pi(\rho)$ is positive, the function can be inverted. That is, the function can be used, in addition to reading off the smallest no-fit rate that is congruent with a certain no-observation rate, for the given data and model, to read off the smallest no-observation rate that needs to be assumed to achieve a certain no-fit rate. Therefore the tabular or graphical form of the $\pi(\rho)$ function is a diagnostic tool to judge model fit with missing data taken into account.

In the remainder of this section, the use of $\pi(\rho)$ will be illustrated for model diagnostics. The data we use here is a cross classification of US respondents from ISSP (1995) according to how proud they are of the way democracy works in their country and of achievements in sport. Only respondents who answered both of these questions will be taken into account now. Note that our framework may be modified to include information from those who answered one question only. In the data, there are 1219 complete observations (i.e., observations for both variables). The categories of both variables are very proud, somewhat proud, not very proud, not proud at all. The model investigated here is independence of the two variables. For the model, the Pearson chi squared statistics is 114.52 with 9 degrees of freedom, indicating that the data provide the analyst with quite strong evidence against independence (for the entire population). The value of the mixture index of fit is 13.59%, that is, based on the data, one estimates at least 13.59% of the population to be outside of independence.

The values of $\pi(\rho)$ can be computed for various values of ρ. These computations are based on the observed distribution *O* and the values obtained for $\pi(\rho)$ can be considered as estimates of its theoretical counterpart, that is, of a similar function, defined in terms of the true distribution on the part of the population that is available for observation, just like *O* can be considered an estimate of this true distribution. The function $\pi(\rho)$ will reach 0 for a certain value of ρ and remains zero for other ρ values larger than this one. In the range of ρ values where the function is positive, it is strictly monotone decreasing. This property makes it possible to apply interpolation for inference concerning ρ values for which the function was not actually computed. Figure 1 shows the $\pi(\rho)$ function based on calculations for selected values of ρ. The behavior of $\pi(\rho)$ is as expected from the theoretical considerations. It is strictly monotone decreasing before it reaches zero and then remains constant. It reaches zero at $\rho = 28.88\%$, that is, if one assumes a no-observation rate of 28.88%, it is possible that independence is valid for the entire population. For smaller no-observation rates, there is a positive no-fit rate. For example, assuming a 10% no-observation rate, independence may account for as much as 91.7% of the population (i. e., the no-fit rate is 8.3%). Or, by assuming a 5% no-observation rate, the no-fit rate is estimated to be 10.4%. Here, the assessment of model fit depends on how realistic a 5% or 10% no-observation rate and the assumed distribution of those who were not available for observation and, further, how satisfactory a fit rate of around 90% appears to the analyst. Notice, that if a representation of the form (3) with a certain ρ and π is possible, a representation with the same π but a larger ρ is also possible. Therefore, the analyst does not have to assess how realistic a certain no-observation rate ρ is rather, how realistic is the assumption, that the no-observation rate is at least ρ. There are very few surveys in practice where the assumption of a no-fit rate of at least 5% or 10% is unrealistic.

*** insert Figure 1 around here ***

Table 7 shows a representation of the form (3) for the ISSP data with ρ=5%. In this case, the no-fit rate is little over 10%. Most of the missing observations (more than 85% of them), belong to the cell very proud of democracy and somewhat proud of sport and the part for the population where independence is not true, is mostly characterized (over 86%) by the cell which is very proud of both democracy and sport. Therefore, the deviation from independence in the data is characterized by too few people who are very proud of democracy and somewhat proud of sport and too many people who are very proud of both. In fact, these two cells have the largest contribution to the Pearson chi-squared of 114,52. Their total contribution is 65.51, leaving 49.01 to the remaining 14 cells.

*** insert Table 7 around here ***

All the diagnostic results discussed so far were actually computed. The $\pi(\rho)$ function in Figure1 however, may be used to approximate values that were not computed. For example, the analyst may take the position that a statistical model is only deemed relevant for the underlying population if it may be able to describe as much as 95% of it. From Figure 1, one finds that $\pi(\rho)=0.05$ for $\rho=0.16$, approximately. That is, one has to assume a no-observation rate of 16% to be able to assume that the model may describe 95% of the population.

4. CONCLUSIONS AND FURTHER RESEARCH

The approach outlined in this paper may serve as a replacement or complement to other methods of assessing model fit, when the researcher wishes to take the potential effect of missing data into account. This approach is based on a non restrictive framework and assumes that a certain part of the population was not available for observation and, consequently, the available data describe only a fraction of the population. The observed data are augmented, using a mixture representation, with hypothetical data for the unobserved part of the population. Then, the mixture index of fit is applied to the augmented (mixture) data to assess model fit. The most important inferential procedure is to estimate, for a given no-observation rate ρ, the smallest fraction $\pi(\rho)$ of the population that cannot be described by the model. The

$\pi(\rho)$ function has intuitively appealing properties. Assessment of model fit may be based either on the value of $\pi(\rho)$, for a known or realistic value of ρ, or on the value of ρ that has to be assumed to achieve a desirably low value of π.

When the potential effect of missing data is also taken into account, the data, typically, provide the researcher with less evidence against the structural models than if standard statistical techniques are used without taking the missing data into account. This fact may come as good news for those who feel there is a strong contradiction between the approximate nature of the available data sets (in the sense that they, at most, approximate the true distribution but because of missing information never really represent it), on the one hand, and the strict procedures proposed by mathematical statistics. For many researcher, this contradiction presents itself in the common experience that when one has large data sets (which is very desirable because of other considerations), simple models usually fail to show acceptable fit. The approach outlined in this paper is not sensitive to the sample size in the traditional sense. Moreover, the method of determining confidence bounds for the mixture index of fit, described in Rudas, Clogg, Lindsay (1994), apply directly to the computation of confidence bounds for $\pi(\rho)$, in the case of fixed ρ.

As it was mentioned earlier, the framework may be extended to include item nonresponse. This requires adding further components to the mixture representation on the left hand side of (3) according to the different patterns of missing observation. By appropriate estimation techniques, the standard assumptions regarding the missing data mechanism (see Little, Rubin, 1987) may also be incorporated.

In theory, the approach presented here can be applied to contingency tables of any size and to any statistical model. However, the presently available estimation methods (Verdes, 2002), although they work for larger problems, tend to perform poorly. This is mostly due to the fact that the optimization problem of determining $\pi(\rho)$ has several local solutions and to find a global optimum, one has to start the algorithm with a large number of different starting values. In addition to algorithmic developments, the extension of the framework to continuous data, like the extensions of the mixture index of fit in Rudas, Clogg, Lindsay (1994), in Rudas (1998, 1999) and Knott (2001) will be investigated in the future.

ACKNOWLEDGEMENTS

The research reported in this paper was supported in part by Grant T-032213 from the Hungarian National Science Foundation (OTKA). The author is indebted to Emese Verdes for programming and carrying out the computations of the examples presented and to the Zentralarchiv fur Empirische Sozialforschung (Cologne) for supplying the ISSP (1995) data..

REFERENCES

Blau, P. M., Duncan, O. D. (1967) *The American Occupational Structure.* New York: Free Press.

Clogg, C. C., Rudas, T., Xi, L. (1995) A new index of structure for the analysis of models for mobility tables and other cross classifications. *Sociological Methodology* 25: 197-222.

Clogg, C. C., Rudas, T., Matthews, S. (1998) Analysis of model misfit, structure, and local structure in contingency tables using graphical displays based on the mixture index of fit. in: Greenacre M., Blasius, J. (eds.) *Visualization of Categorical Data,* 425-439. San Diego: Academic Press.

Dayton . M (2003) Applications and Computational Strategies for the Two-Point Mixture Index of fit. *British Journal of Mathematical & Statistical Pszchology*. In press.

ISSP (1995) See: http://www.gesis.org/en/data_service/issp/index.htm

Formann, A. K. (2000) Rater agreement and the generalized Rudas-Clogg-Lindsay index of fit. *Statistics in Medicine*. 19, 1881-1888.

Formann, A. K. (2003) Latent class model diagnostics – a review and some proposals. *Computational Statistics and Data Analysis. 19, 549-559*

Knoke, D., Burke, P. J. (1980) *Log-Linear Models.* Newbury Park: Sage.

Knott, M. (2001) A measure of independence for a multivariate normal distribution and some connections with factor analysis. *Research Report 63*, Department of Statistics, London School of Economics.

Little, R. J. A., Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

MATLAB (2002) See: http://www.mathworks.com/products/matlab

Rudas, T. (1998) Minimum mixture estimation and regression analysis. in: Marx, B., Friedl, H. (eds). *Proceedings of the 13th International Workshop on Statistical Modeling*, 340-347, Louisiana State University.

Rudas, T. (1999) The mixture index of fit and minimax regression. *Metrika*, 50: 163-172.

Rudas, T. (2002) A latent class approach to measuring the fit of a statistical model. in: Hagenaars, H., McCutcheon, A. (eds.) *Applied Latent Class Analysis*. Cambridge: Cambridge University Press.

Rudas, T., Clogg, C. C., Lindsay, B. G. (1994) A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Ser. B,* 56: 623-39.

Rudas, T., Zwick, R. (1997) Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, 22: 31-45.

Verdes, E. (200) See: http://www.klte.hu/~vemese/spistar.htm

Xi, L. (1996) Measuring goodness-of-fit in the analysis of contingency tables with mixture based indices: algorithms, asymptotics and inference. *Ph. D. Thesis*, Department of Statistics, The Pennsylvania State University.

Xi, L., Lindsay, B. G. (1996) A note on calculating the π* index of fit for the analysis of contingency tables. *Sociological Methods and Research,* 25: 248-259.

Table 1

A hypothetical observed distribution

0.1	0.3
0.2	0.4

Table 2

Mixture representation of the distribution in Table 1

Observed

0.8 x

0.1	0.3
0.2	0.4

+

Not observed

0.2 x

0.2	0.0
0.2	0.6

=

Fit

0.93 x

0.129	0.258
0.2043	0.4086

+

Unobserved

0.07 x

0.0	0.0
0.1429	0.857

Table 3

Estimated distribution of the population, based on Table 2

0.12	0.24
0.20	0.44

Table 4

Another mixture representation of the distribution in Table 1.

Observed

0.9 x

0.1	0.3
0.2	0.4

\+

Unobserved

0.1 x

0.6	0.4
0.0	0.0

=

Fit

0.99 x

0.1515	0.3030
0.1818	0.3636

\+

No fit

0.01 x

0.0	1.0
0.0	0.0

Table 5

The fit of three statistical models to the Blau-Duncan data

Model	Degrees of freedom	Pearson chi-squared	Likelihood ratio	Mixture index of fit	$\pi(\rho)$		
					5%	10%	15%
Independence	16	875.10	830.98	0.310	0.205	0.166	0.127
Quasi-independence	11	209.07	255.14	0.147	0.072	0.058	0.039
Quasi-uniform association	10	30.78	27.82	0.052	0.010	0.005	0.001

Table 6

Representation of the Blau-Duncan data as in (3)

(Probabilities multiplied by 100)

Observed

0.900 x

4.4.8	1.94	0.97	1.15	0.12
5.92	4.68	2.12	2.36	0.24
4.06	3.68	5.42	5.06	0.21
4.21	4.74	6.15	11.13	0.50
2.89	4.30	6.10	10.92	6.65

+

Unobserved

0.100 x

0.00	0.00	5.16	14.59	0.00
0.00	0.00	5.84	23.96	0.00
0.11	0.02	0.00	38.72	0.00
0.00	0.00	0.03	0.00	0.00
11.56	0.00	0.00	0.00	0.00

=

Fit

0.944 x

4.27	1.04	1.47	2.64	0.06
1.82	4.46	2.64	4.78	0.11
3.40	3.51	5.17	8.92	0.20
4.01	4.14	5.87	10.61	0.23
3.97	4.10	5.81	10.42	6.34

+

No fit

0.056 x

0.00	13.76	0.00	0.00	0.92
64.52	0.00	0.00	0.00	2.02
8.20	0.00	0.00	0.02	0.00
0.00	6.41	0.00	0.00	4.15
0.00	0.00	0.00	0.00	0.00

Table 7

A representation of the ISSP (1995) data with ρ=5%

(probabilities multiplied by 100)

Observed

0.95 x

		Sport			
		Very proud	Somewhat proud	Not very proud	Not proud at all
Democracy	Very proud	17.23	9.43	1.89	0.57
	Somewhat proud	16.49	31.91	4.35	1.72
	Not very proud	3.53	7.71	1.72	0.9
	Not proud at all	0.57	1.31	0.33	0.33

+

Unobserved

0.05 x

		Sport			
		Very proud	Somewhat proud	Not very proud	Not proud at all
Democracy	Very proud	0	85.55	0.2	3.4
	Somewhat proud	0.07	0.01	0.01	0.04
	Not very proud	8.72	0	0	0.01
	Not proud at all	1.98	0	0	0

=

Fit

0.8961 x

		Sport			
		Very proud	Somewhat proud	Not very proud	Not proud at all
Democracy	Very proud	7.68	14.87	2.03	0.8
	Somewhat proud	17.61	34.07	4.64	1.84
	Not very proud	4.26	8.23	1.12	0.44
	Not proud at all	0.72	1.4	0.19	0.08

+

No fit

0.1039 x

		Sport			
		Very proud	Somewhat proud	Not very proud	Not proud at all
Democracy	Very proud	86.5	0	0	0
	Somewhat proud	0	0	0	0
	Not very proud	0	0	5.8	4.19
	Not proud at all	0	0	1.29	2.22

Figure 1.

The $\pi(\rho)$ function for the ISSP data

(Fractions multiplied by 100)