Wu Chou, etc.

CRC Book

CRC PRESS
Boca Raton Ann Arbor London Tokyo

Contributors

Dobra
EroshevaFienberg

0-8493-0052-5/00/$0.00+$.50

Contents

0-8493-0052-5/00/$0.00+$.50

1

Disclosure Limitation Methods Based on Bounds for Large Contingency Tables With Applications to Disability

Adrian Dobra, Elena A. Erosheva and Stephen E. Fienberg
Duke University, Durham and Carnegie Mellon University, Pittsburgh

CONTENTS

Much attention has been focused recently on the problem of maintaining the confidentiality of statistical data bases through the application of statistical tools to limit the identification of information on individuals (and enterprises). Here we describe and implement some simple procedures for disclosure limitation based on bounds for the cell entries in contingency tables that result from knowledge about released marginal totals or subtables. Our work draws on the ideas associated with the theory of log-linear models for contingency tables where the minimal sufficient statistics are in fact marginal totals corresponding to the highest-order terms in the model. We draw on recent results associated with decomposable log-linear models and their use in the disclosure limitation context.

Our primary illustration of the methodology is in the context of a 2^{16} contingency table extracted from disability data collected as part of the National Long Term Care Survey. We treat these data as if they involved an entire population and we illustrate the calculation of optimal releases of marginals in such a circumstance. We describe briefly some of the analyses we have carried out on these data using the Grade of Membership model, whose minimal sufficient statistics are not simply marginal totals, and we relate this to the optimal set of releasable margins. We conclude with a discussion of some of the possible implications of our analyses for disclosure limitation in similar data sets.

0-8493-0052-5/00/$0.00+$.50

1.1 Introduction

If government agencies are to collect and publish high quality data, it is essential that they maintain the confidentiality of the information provided by others. Typically agencies promise respondents that their data will be kept confidential and used for statistical purposes only. Disclosure limitation is the process of protecting the confidentiality of statistical data. A disclosure occurs when someone can use published statistical information to identify an individual data provider. Since virtually any form of data release contains some information about the individuals whose data are included in it, disclosure is not an all-or-none concept but rather a probabilistic one. For general introductions to some of the statistical aspects of confidentiality and disclosure limitation see Doyle, et al. [13], Duncan, et al. [15], Fienberg [20], and Willenborg and De Waal [37, 38].

Disclosure limitation procedures alter or limit the data to be released, e.g., by modifying or removing those characteristics that put confidential information at risk for disclosure. In the case of sample categorical data in the form of a contingency table, a count of "1" can generate confidentiality concerns if that individual is also unique in the population. Much confidentiality research has focused on measures of risk that attempt to infer the probability that an individual is unique in the population given uniqueness in the sample (e.g, see Chen and Keller-McNulty [4], Fienberg and Makov [22, 23], Skinner and Holmes [36], and Samuels [35]). Here we will consider only the case of population data, for which a count of "1" is unique. Moreover, a count of "2" is also problematic for population data since it allows each of the two people in the cell to identify one other! More generally, small counts raise issues of disclosure risk.

In this paper we provide an overview of some recent work to develop bounds for entries in contingency and other non-negative tables (see Dobra and Fienberg [9, 10, 11], and Dobra, et al. [12]). We work within a statistical framework for the release of cross-classified categorical data coming originally in the form of a contingency table where requests from users come in the form of (marginal) subtables involving a subset of the variables. Clearly, the more such subtables that are available, the more information we have about the full joint distribution of the cross-classifying variables. Through a detailed example we illustrate both the utility of data releases in the form of marginals and simple methods for assessing the risk of disclosure using bounds on the individual cell entries. Our interest in this problem grows out of work to develop a Web-based table query system, coordinated by the National Institute of Statistical Sciences [12]. The system is being designed to work with a database consisting of a k-way contingency table and it allows only those queries that come in the form of requests for marginal tables. What is intuitively clear from statistical theory is that, as margins are released and cumulated by users, there is increasing information available about the table entries. Such an approach to disclosure limitation always tell the truth by releasing marginals from the full table, albeit not the whole truth, which would entail releasing the full table (c.f., Dobra, et

al. [12]).

The approach we outline in this paper draws heavily on the ideas associated with the theory of log-linear models for contingency tables (see Bishop, Fienberg, and Holland [1], and Lauritzen [31]), where the minimal sufficient statistics (MSSs) are in fact marginal totals corresponding to the highest-order terms in the model. This simple statistical fact has profound implications for reporting purposes as well as for disclosure limitation methods based on reporting only subtables. If an agency knows that a particular log-linear model fits a multi-dimensional contingency table well, then, at least in principle, users of the data could get by with only the MSSs. If the agency is able to release a set of marginals which include the MSSs of well fitting log-linear models, then users can also independently assess the fit relevant log-linear models from the released data and consider alternative models as well. It is in this sense that an approach based on releasing marginals leads to conclusions that may be more uncertain, but will not be erroneous.

In the next section, we introduce an example of a 2^{16} contingency table based on disability data from the National Long Term Care Survey, which we use to illustrate our methods. In Section 1.3 we give a brief summary of the key technical background on bounds for cell entries in a table when the marginals corresponding to those associated with decomposable and reducible graphical models. Then, in Sections 1.4 and 1.5, we outline a general approach to the determination of optimal releases of marginals based on a search procedure that involves only decomposable cases and apply it to our example. In Section 1.6 we assess our results to the 2^{16} table, and we conclude with a discussion of some of the possible implications for disclosure and statistical analyses.

1.2 Example: National Long Term Care Survey Data

In this paper our primary example is a 2^{16} contingency table **n** extracted from the "analytic" data file for National Long-Term Care Survey. Each dimension corresponds to a measure of disability defined by an activity of daily living, and the table contains information cross-classifying individuals aged 65 and above. This extract involves data pooled across four waves of a longitudinal survey, and it involves sample as opposed to population data. We henceforth act *as if* these were population data. For a detailed description of this extract see [17].

The 16 dimensions of the contingency table correspond to responses to 6 activities of daily living (ADLs) and 10 instrumental activities of daily living (IADLs). Specifically, the ADLs are (1) *eating*, (2) *getting in/out of bed*, (3) *getting around inside*, (4) *dressing*, (5) *bathing*, (6) *getting to the bathroom or using a toilet*. The IADLs are (7) *doing heavy house work*, (8) *doing light house work*, (9) *doing laundry*, (10) *cooking*, (11) *grocery shopping*, (12) *getting about outside*, (13) *traveling*, (14) *managing money*, (15) *taking medicine*, (16) *telephoning*. For each ADL/IADL

measure, subjects were classified as being either disabled (level 1) or healthy (level 0) on that measure.

Of the $2^{16} = 65,536$ cells in the table, 62,384 (95.19%) contain zero entries, $1,729$ (2.64%) contain counts of "1", 499 (0.76%) contain counts of "2". The largest cell count is $3,853$, in the $(0,0,\ldots,0)$ cell corresponding to being healthy on all 16 measures. In fact, no relatively simple hierarchical log-linear model provides a reasonable fit to these data in part because they all substantially underestimate the value of this cell count in particular.

In the absence of simple parsimonious log-linear models to describe such disability data, considerable attention has been given to analyses using what is known as the Grade of Membership (GoM) model (see Manton, Woodbury, and Tolley [33] and [19]). The GoM model is a partial or mixed membership model that resembles a more traditional latent class model. For a random sample of subjects, we observe K dichotomous responses, $x_1,\ldots,x_K$. We assume there are J basis subpopulations, which are determined by the conditional (positive) response probabilities, $\lambda_{jk},\ k = 1,\ldots,K$. The subjects are characterized by their degrees of membership in each of the subpopulations, $g = (g_1,\ldots,g_J)$, which are nonnegative and add to 1. Conditional on the subject's membership scores, g, the subject's response probability for item k is given by a convex combination $\Pr(x_k = 1|g) = \sum_j g_j\lambda_{jk}$. We assume that the responses $x_1,\ldots,x_K$ are conditionally independent, given the membership scores. For many purposes we may also want to add the assumption that the membership scores, g, have a Dirichlet distribution with parameters $\alpha = (\alpha_1,\ldots,\alpha_J)$. For the disability data in our example, $K = 16$ and a "reasonable" value of $J = 5$ (e.g., see Erosheva [17, 18]).

The GoM likelihood function is not of the exponential family type, and thus no sufficient statistics exist for the membership scores [17]. This does not allow for conditional likelihood estimation and also means that if the GoM model is an appropriate one to describe the disability data in the 2^{16} table, then we need more than the simple marginal totals associated with any unsaturated log-linear model to estimate the GoM parameters. We return to this point after we explore the disclosure limitation properties of bounds based on the release of marginal tables for these data.

1.3 Technical Background on Cell Entry Bounds

Bounds for entries in two-way contingency tables go back to seminal papers by Bonferonni [2], Fréchet [26], and Hoeffding [27]. For an $I \times J$ table with entries $\{n_{ij}\}$ and row margins $\{n_{i+}\}$ and column margins $\{n_{+j}\}$, these bounds take the form

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}. \tag{1.1}$$

For simplicity, we refer to these as *Fréchet bounds*. Until recently, the only multidimensional generalizations of this result that have been utilized involved non-overlapping

FIGURE 1.1
Independence graph for a 6-dimensional table and a log-linear model induced by the marginals [BF], [ABCE] and [ADE]

fixed marginals (c.f. the related work described in Joe [28]).

Any contingency table with non-negative integer entries and fixed marginal totals is a lattice point in the convex polytope **Q** defined by the linear system of equations induced by the released marginals. The constraints given by the values in the released marginals induce upper and lower bounds on the interior cells of the initial table. These bounds or *feasibility intervals* can be obtained by solving the corresponding linear programming problems. The importance of systematically investigating these linear systems of equations should be readily apparent. If the number of lattice points in **Q** is below a certain threshold, we have significant evidence that a potential disclosure of the entire dataset might have occurred. Moreover, if the induced upper and lower bounds are too tight or too close to the actual sensitive value in a cell entry, the information associated with the individuals classified in that cell may become public knowledge.

The problem of determining sharp upper and lower bounds for the cell entries subject to some linear constraints expressed in this form is known to be NP-hard (see Roehrig et al. [34]). Several approaches have been proposed for computing bounds: however, almost all of them have drawbacks that show the need for alternate solutions.

We visualize the dependency patterns induced by the released marginals by constructing an independence graph for the variables in the underlying cross-classification. Each variable cross-classified in the table is associated with a vertex in this graph. If two variables are not connected, they are conditionally independent given the remaining variables. Models described solely in terms of such conditional independencies are said to be *graphical* (e.g. see Lauritzen [31]). For example, Figure 1.1 shows the independence graph for a 6-variable cross-classification with the variables $\{A,B,C,D,E,F\}$ corresponding to the 6 nodes. Of the 15 possible edges, 6 are absent and correspond to conditional independencies.

Decomposable models are a subclass of graphical models that correspond to triangulated graphs and have closed form structure and special properties. In particular, the expected cell values can be expressed as a function of the fixed marginals. To

be more explicit, the maximum likelihood estimates are the product of the marginals divided by the product of the separators. For example, the graph in Figure 1.1 is triangulated and, for the corresponding decomposable log-liner model, the marginals [BF], [ABCE], and [ADE], corresponding to the cliques in the graph, are the MSSs. The cliques are "separated" from one another by subsets of connected nodes, which we refer to as separators.

By induction on the number of MSSs, Dobra and Fienberg [9], developed generalized Fréchet bounds for sets of margins that correspond to the MSSs of any decomposable log-linear model. These generalized Fréchet bounds are sharp in the sense that they are the tightest possible bounds given the marginals and there are feasible tables for which these bounds are attained.

Theorem 1 (Fréchet Bounds for Decomposable Models). *Assume that the released set of marginals for a K-way contingency table correspond to the MSSs of a decomposable log-linear model. Then the upper bounds for the cell entries in the initial table are the minimum of relevant margins, while the lower bounds are the maximum of zero, or sum of the relevant margins minus the separators.*

When the log-linear model associated with the released set of marginals is not decomposable, it is natural to ask ourselves whether we could reduce the computational effort needed to determine the tightest bounds by employing the same strategy used for decomposable graphs, i.e. decompositions of graphs by means of complete separators. An independence graph that is not necessarily decomposable, but still admits a *proper* decomposition (i.e., looks like a decomposable graph but whose components are not fully connected), is called *reducible* (Leimer [32]). Once again, we point out the link with maximum likelihood estimation in log-linear models. We define a *reducible log-linear model* in [9] as one for which the corresponding MSSs are marginals that characterize the components of a reducible independence graph. If we can calculate the maximum likelihood estimates for the log-linear models corresponding to every component of a reducible graph $\mathscr{G}$, then we can easily derive explicit formulae for the maximum likelihood estimates in the reducible log-linear model with independence graph $\mathscr{G}$ [9].

Theorem 2 (Fréchet Bounds for Reducible Models). *Assume that the released set of marginals is the set of MSSs of a reducible log-linear model. Then the upper bounds for the cell entries in the initial table are the minimum of upper bounds of relevant components, while the lower bounds are the maximum of zero, or sum of the lower bounds of relevant components minus the separators.*

Finally, we note that when the released margins correspond to a log-linear model that is neither decomposable nor reducible, a more elaborate form of bounds calculation is required. Dobra [7, 11] has developed an iterative algorithm for this situation, generalizing the original "shuttle" procedure proposed by Buzzigoli and Giusti [3], which can be used to compute sharp bounds. Unfortunately, as the dimensionality of the table grows, this algorithm is computationally elaborate and is not especially useful as the main component of a search for an optimal form of marginal release.

Instead we adopt simplified search strategies and then use the algorithm only after we have focused in on a small subset of sets of marginal releases. In the following sections, we turn to such simplified search strategies that exploit the bounds calculation in the decomposable case. When we apply these to the 2^{16} table from Section 1.2, we also describe the results of applying the generalized shuttle algorithm.

1.4 Decomposable Frontiers

Here we briefly describe the method of Dobra et al. [12] to identify a releasable set of marginals based on a search using decomposable bounds and we apply the approach to out 2^{16} example.

The set **S** of all 2^k marginals of a k-way table **n** is partially ordered by set inclusion of variable. If the variables associated with a marginal $\mathbf{n}_1$ are contained in the set of variables associated with another marginal $\mathbf{n}_2$, we say that $\mathbf{n}_1$ is a *child* of $\mathbf{n}_2$ and $\mathbf{n}_2$ is a *parent* of $\mathbf{n}_1$. The *released frontier* $\mathscr{RF}$ of a set $\mathscr{R}$ of released marginals consists of the maximal elements of $\mathscr{R}$–those with no released parents. Clearly, any set of released marginals is completely identified by its frontier. The elements of $\mathscr{RF}$ consist of sets of marginals and they represent the trade-offs that occur when the release of some marginals make others unreleasable. Our goal here is to identify useful elements of $\mathscr{RF}$.

For simplicity, we consider a set $\mathscr{R}$ to be releasable if and only if the minimum difference between the upper and lower bounds for the small count cells of "1" or "2" in table **n** is greater or equal to some threshold β.

In the context of the Web-based query system, the set **S** is partitioned at any time t as follows:

$$\mathbf{S} = \mathscr{R}(t) \cup \mathscr{M}(t) \cup \mathscr{U}(t), \tag{1.2}$$

where $\mathscr{R}(t)$ are the released marginals at time t, $\mathscr{M}(t)$ are the possible future releases at time t, and $\mathscr{U}(t)$ are the marginals that became un-releasable by releasing $\mathscr{R}(t)$. As we release additional marginals, we select elements from $\mathscr{M}(t)$ for inclusion in $\mathscr{R}(t)$ and at the same time move other elements into $\mathscr{U}(t)$.

W may not want to allow all elements in $\mathscr{U}(t)$ to be a potential release at time t because the release of some would essentially foreclose on the possibility of releasing others at a later time. Therefore, the system might also maintain a list of candidate releases $\mathscr{CM}(t) \subset \mathscr{M}(t)$.

Now assume a user requests a marginal $\mathbf{n}_0$. In order to accept or deny this request, the system would have to dynamically evaluate whether the set $\mathscr{R}(t) \cup \{\mathbf{n}_0\}$ is releasable provided that $\mathbf{n}_0$ belongs to $\mathscr{CM}(t)$. If $\mathbf{n}_0$ is released, the system needs to update the sets $\mathscr{U}(t+1)$ and $\mathscr{CM}(t+1)$ very quickly to be ready to process a new request. In addition, evaluating the disclosure risk is a lot more difficult if the system

takes into account the fact that it gives away information about **n** when denying a request.

In actual applications, the underlying categorical database **n** might have 40 or more dimensions and/or millions and millions of cells. Consequently, dynamically evaluating whether a marginal is releasable as well us updating the sets $\mathscr{U}$ and $\mathscr{C}.\mathscr{M}$ involve huge computations that cannot be done on today's computers. Besides scalability issues, there are other concerns relating to user equity: if we release a marginal, some other marginals become unreleasable and hence those users requesting these marginals might suffer if a policy of "first come, first served" would be applied.

A possible solution would be to replace sequential releases with one-time releases. In this case, the complete set of marginal **S** contains the released marginals $\mathscr{R}$ and the un-released marginals $\mathscr{U}$. The only difficulty of this approach is identifying the "best" $\mathscr{R}$ according to some data utility criteria. The tedious dynamic risk computations are now replaced by a one-time computation that can be done offline. Users can be polled on the choice of $\mathscr{R}$. This simplified static version of the table server is not prone to be attacked by intruders as the dynamic server was.

1.4.1 Calculating Decomposable Frontiers

we say that a release $\mathscr{R}$ is *decomposable* if its corresponding frontier defines the MSSs of a decomposable graphical model. A decomposable frontier is the frontier of a decomposable release. In this case, the upper and lower bounds induced by $\mathscr{R}$ can be computed using formulas [9], which reduces to almost zero the computational effort required to establish whether $\mathscr{R}$ is releasable or not.

We quantify the data utility $\mathbf{DU}(\mathscr{R})$ of a release $\mathscr{R}$ by the total number of marginals contained in $\mathscr{R}$. To maximize $\mathbf{DU}(\mathscr{R})$ over the space of decomposable releasable sets $\mathscr{R}$ we use a simulated annealing approach that involves generating random draws from the distribution

$$\pi(\mathscr{R}) \propto \exp(\mathbf{DU}(\mathscr{R})/T), \tag{1.3}$$

where T is a scale parameter called *temperature* The temperature T is slowly decreased toward 0 as the algorithm progresses. Given a current state $\mathscr{R}_0$, a new decomposable set of sub-tables $\mathscr{R}_1$ is selected from a uniform distribution on a neighborhood $N(\mathscr{R}_0)$ of $\mathscr{R}_0$. If $\mathbf{DU}(\mathscr{R}_1) \geq \mathbf{DU}(\mathscr{R}_0)$, $\mathscr{R}_1$ is "accepted" with probability 1, that is $\mathscr{R}_1$ becomes the current state. Otherwise, if $\mathbf{DU}(\mathscr{R}_1) < \mathbf{DU}(\mathscr{R}_0)$, $\mathscr{R}_1$ could be accepted with probability

$$\min\{\exp((\mathbf{DU}(\mathscr{R}_1) - \mathbf{DU}(\mathscr{R}_0))/T), 1\}. \tag{1.4}$$

We repeat this simulation process and the resulting sequence The Markov chain $\{\mathscr{R}_j\}$ forms a Markov chain that will concentrate in a smaller and smaller region around a local maxima of $\mathbf{DU}(\mathscr{R})$ as T approaches 0. Therefore, at higher temperatures, the simulated annealing algorithm can "escape" local optima of the criterion function and eventually converge to a global optimum.

The neighborhood $N(\mathscr{R})$ of a decomposable set of sub-tables $\mathscr{R}$ is taken to be all the sets of sub-tables determined by decomposable independence graphs obtained by deleting or adding one edge from the independence graph associated with $\mathscr{R}$. Very efficient algorithms for finding $N(\mathscr{R})$ are presented in [6]. Any two decomposable graphs can be "linked" by a sequence of decomposable graphs that differ by exactly one edge (see, for example, [31]), and hence the resulting Markov chain is irreducible.

1.4.2 Analysis of the 2^{16} NLTCS Example

It is standard survey practice to release the one-way marginals for all variables, and thus we begin by assuming that these have already been released. We ran the simulated annealing algorithm for searching a decomposable frontier for three different threshold values, $\beta = 3, 4, 5$. The resulting decomposable frontiers are:

$$\begin{aligned}\mathscr{RF}(\beta=3) = \Big\{&[5,10,12,13,14,15,16],[5,10,11,14,15,16],[9,10,12,13,14,15],\\ &[6,10,12,13,15,16],[4,10,12,13,14,15],[4,8,10,12,13,14],\\ &[3,4,12,13,14,15],[3,4,7,12,13,15],[2,12,13,14,15,16],\\ &[1,9,12,13,14,15]\Big\},\end{aligned} \tag{1.5}$$

$$\begin{aligned}\mathscr{RF}(\beta=4) = \Big\{&[6,9,12,13,15,16],[6,8,12,13,15,16],[2,6,8,12,13,15],\\ &[2,6,11,12,13,15],[2,4,6,11,12,13],[2,4,11,12,13,14],\\ &[2,4,6,10,12,13],[2,4,5,10,12,13],[2,3,6,8,12,13],\\ &[1,8,12,13,15,16],[2,4,6,7,11]\Big\},\end{aligned} \tag{1.6}$$

$$\begin{aligned}\mathscr{RF}(\beta=5) = \Big\{&[6,8,10,14,15,16],[4,6,8,10,14,15],[15,14,8,6,4,3],\\ &[3,4,6,8,13,15],[3,4,6,12,14,15],[3,4,6,9,13,15],\\ &[2,4,6,8,13,15,13],[2,4,8,11,13,15],[3,4,6,7,14],\\ &[4,5,6,12,14,12],[1,4,6,14,15]\Big\},\end{aligned} \tag{1.7}$$

Two of these frontiers contain 6-dimensional marginals, while the third, $\mathscr{RF}(\beta = 3)$, contains a 7-dimensional marginal. Summaries of the released sets of marginals determined by these frontiers are presented in Tables 1.1, 1.2 and 1.3.

The releasable frontier consists of *multiple* sets of releases, and for each the released marginals are maximal in the sense that any additional marginal is unreleasable. The simulated annealing algorithm happened to find the sets of releases on the releasable frontiers given above, but there are likely several other frontier elements.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	66	120	55%
3	125	560	22.32%
4	125	1,820	6.87%
5	66	4,368	1.51%
6	16	8,008	0.20%
7	1	11,440	0.00%

Breakdown of the released set of sub-tables $\mathscr{RF}(\beta = 3)$. The columns show the dimension of sub-tables, how many sub-tables of that dimension are in $\mathscr{RF}(\beta = 3)$, the total number of sub-tables and the percentage of released sub-tables. The total number of released sub-tables is 415.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	64	120	53.33%
3	116	560	20.71%
4	109	1,820	5.99%
5	52	4,368	1.19%
6	10	8,008	0.12%

Breakdown of the released set of sub-tables $\mathscr{RF}(\beta = 4)$. The total number of released sub-tables is 367.

As the threshold β decreases, the number of released sub-tables increases for each dimension. Examining the decomposable frontiers for $\beta = 3, 4, 5$, we first notice that they are not nested. For example, the released sets of sub-tables defined by frontiers $\mathscr{RF}(\beta = 4)$ and $\mathscr{RF}(\beta = 5)$ are not subsets of the set of sub-tables defined by $\mathscr{RF}(\beta = 3)$. Note that all marginals of the "most generous" decomposable frontier $\mathscr{RF}(\beta = 3)$ contain 0-2 ADL and 4-6 IADL variables, but most marginals of the frontier $\mathscr{RF}(\beta = 5)$ contain 3 ADL and 3 IADL variables. Thus, it seems that releasing fewer ADL variables in the marginals allows us to maximize the total number of marginals released for a lower value of threshold β. This might be related to an existing theory which says that ADL variables are approximately hierarchical, e.g., see Katz et al. [29]. If the small counts of 1 and 2 are indicative of "imbalance" in the marginals, and if responses on ADL items are more structured than responses on IADL items, releasing more IADL items is "safer" than releasing more ADL items.

1.5 "Greedy" Frontiers

Searching for decomposable releases, although appealing from a computational point of view, can be considered to be too restrictive from a practical perspective. In this

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	62	120	51.67%
3	108	560	19.29%
4	97	1,820	5.33%
5	44	4,368	1.00%
6	8	8,008	0.10%

Breakdown of the released set of sub-tables $\mathscr{RF}(\beta = 5)$. The total number of released sub-tables is 335.

section, we present a heuristic procedure for identifying an arbitrary release that is based on a consistent methodology for assessing the disclosure risk associated with releasing a particular marginal. We illustrate the components of this algorithm using the 2^{16} table example.

We begin by introducing the notion of the *most parsimonious* model corresponding to a sub-table. Let $K = \{1, 2, \ldots, k\}$ denote the indices of the variables cross-classified in a k-dimensional table $\mathbf{n} = \mathbf{n}_K$.

DEFINITION 1.1 *The most parsimonious model associated with the C-marginal of $\mathbf{n}$ is the model with minimal sufficient statistics*

$$\{C\} \cup \left[\bigcup_{j \in K \setminus C} \{\{j\}\} \right]. \tag{1.8}$$

Definition 1.1 says that the most parsimonious model in which a given marginal appears is defined by that marginal and by the one-dimensional marginals corresponding to the variables in the table which do not appear in that marginal. For example, the most parsimonious model corresponding to the $[1,2]$-marginal of a six-way table has minimal sufficient statistics $\{[1,2,3],[4],[5],[6]\}$.

We would like to find a way to quantify how "problematic" the release of a certain marginal might be. In this context, "problematic" means "potentially problematic" because a marginal is released after some other marginals have already been released. The level of how "problematic" a marginal might be is therefore relative to the rest of the marginals and is not an absolute measure that would have a meaning if it would be considered alone. To define such a measure, we propose looking at all the models in which a given marginal is involved, i.e., we consider all possible sets of releases containing that marginal. The most parsimonious model is embedded in all these sets of releases, and, because it has the loosest bounds, it suffices for us to study only this model. If the release of this model is problematic, the release of all other models are problematic as well. On the other hand, if the release of the most parsimonious model is not problematic, one cannot say anything about all the other models that include the marginal.

There is an immediate intuitive interpretation of evaluating the disclosure risk of a marginal based on its most parsimonious model: in order to see how much "damage" this marginal could do, we release this marginal alone along with some minimal information about the variables not contained in this marginal. Another attractive feature of these most parsimonious models is that they are decomposable, hence calculating the upper and lower bounds associated with them is straightforward and can be done by means of explicit formulas.

By employing the notion of parsimonious models, we completely drop the (very) strong decomposability constraint we imposed when we searched for a decomposable frontier. Moreover, we take into account sets of releases that are only required to be hierarchical! This represents the highest level of generality we could hope to achieve. In addition, the size of the space of releases we take into consideration is huge. A search strategy similar to simulated annealing is hopeless if employed on a space of this size!

DEFINITION 1.2 *The* critical width *of a marginal* $\mathbf{n}_C$ *is the minimum of the difference between the upper and lower bounds for the cells containing small counts of "*1*" or "*2*". These bounds are induced by the most parsimonious model associated with* $\mathbf{n}_C$.

The critical width is the minimum of the difference between the relevant bounds for cells with counts of "1" and "2" because all the cells containing small counts in the table have to be protected in order to consider a release to be safe at a given level. If one such small count cell is not adequately protected according to the risk criteria we employ, we consider the entire release to be problematic.

DEFINITION 1.3 *The marginal* $\mathbf{n}_{C_1}$ *is said to be* more problematic *than the marginal* $\mathbf{n}_{C_2}$ *if the critical width of* $\mathbf{n}_{C_1}$ *is smaller than the critical width of* $\mathbf{n}_{C_2}$.

We calculated the critical widths for all the marginals of the 16-dimensional table. The critical widths corresponding to the one-dimensional marginals of a table are equal by definition since they are all calculated based on the same model–complete independence of the variables in that table. Typically agencies attempt to release at least the one-dimensional marginal corresponding to each variable.

In our case, the critical width associated with the one-dimensional marginals is large, i.e., $2,285$. On the other hand, all the 8-dimensional marginals have a critical width of 1. Hence, by releasing only one eight-way sub-table after releasing all the one-way sub-tables, at least one small count cell will be made public. The critical widths for the marginals of dimension $2, 3, \ldots, 7$ are given in Figure 1.2. As the dimension of released subtables increases, the critical widths decrease, tend to be less scattered and gradually cluster around 1.

The most problematic two-way table is, by far, $[7, 8]$ with a critical width of 8. One obvious reason why this marginal is so problematic is the count of 8 in cell

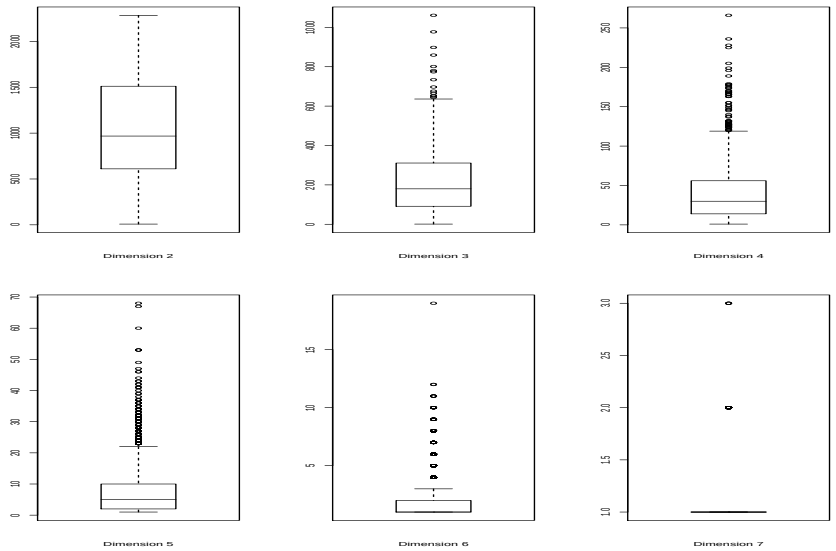

FIGURE 1.2
Boxplots with critical widths associated with marginals of dimension $2,3,\ldots,7$**.**

$(1,0)$ which is very small compared with the other three counts in this marginal. This count corresponds to respondents who could not do IADL *light house work*, but were able to do IADL *heavy house work*. The next most problematic two-way marginal is $[1,7]$ with a critical width of 64, while the third most problematic two-way sub-table is $[5,1]$ with a critical width of 82. The most problematic three-way marginals are $[7,8,12]$, $[7,8,10]$ and $[1,7,8]$, all with a critical width of 3. We note that $[7,8]$ is a child of these three marginals. We also note that the decomposable frontiers of Section 1.4.2 do not contain these most problematic three-way marginals. The most problematic four-way marginals have a critical width of 1. Variables 8, 7 and 1 appear in most of the 36 four-way marginals having this critical width. Other variables, such as 16, 12 and 11 also have a significant presence in these marginals.

To choose a release, we construct a list, $\mathscr{L}$, which contains the marginals in decreasing order with respect to their critical widths. Therefore the least problematic marginals will appear at the top of this list, while the most problematic marginals will be placed at the end. According to our definition, two marginals are "equally problematic" if they have the same critical width. However, to maximize the amount of released information, we might prefer to release, if possible, a higher dimensional marginal instead of a lower dimensional marginal if both marginals have the same critical widths. Consequently, we re-order the marginals in $\mathscr{L}$ having a certain fixed critical width in decreasing order with respect to their dimension.

More explicitly, let $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_2}$ be two marginals with dimensions k_1, k_2 and with critical widths w_1 and w_2. Denote by l_1 and l_2 the ranks of $\mathbf{n}_{C_1}$ and $\mathbf{n}_{C_1}$ in the list $\mathscr{L}$.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	0	16	0%
2	0	120	0%
3	0	560	0%
4	263	1,820	14.45%
5	1,311	4,368	30.01%
6	103	8,008	3.78%

Non-decomposable frontier obtained from the greedy procedure for $\beta = 3$. The total number of sub-tables in this frontier is $1,677$.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	120	120	100%
3	547	560	97.68%
4	1,566	1,820	86.04%
5	1,659	4,368	37.98%
6	103	8,008	3.78%

Breakdown of the released set of sub-tables corresponding to the frontier in Table 1.4. Compare with the released set corresponding with the decomposable frontier $\mathscr{RF}(\beta = 3)$–see Table 1.1. The total number of released sub-tables is $4,011$.

If $w_1 > w_2$, then $l_1 < l_2$. However, if $w_1 = w_2$ and $k_1 > k_2$, then we also require that $l_1 < l_2$.

Let $\mathbf{n}_1$, $\mathbf{n}_2$,..., $\mathbf{n}_L$ be the marginals of $\mathbf{n}$ in the order in which they appear in $\mathscr{L}$. We want to identify the unique rank $l_0 \in \{1,2,\ldots,L\}$ such that the set of marginals $\{\mathbf{n}_1,\ldots,\mathbf{n}_{l_0}\}$ is releasable according to our risk criteria, but $\{\mathbf{n}_1,\ldots,\mathbf{n}_{l_0},\mathbf{n}_{l_0+1}\}$ is not. Instead of sequentially adding new marginals starting from the top of list $\mathscr{L}$, we determine l_0 by employing a much more efficient bisection search strategy.

We used this greedy algorithm to determine a releasable set of marginals for the 16-dimensional table for thresholds 3, 4 and 5. These releases are summarized below. We note that the releases obtained from the greedy algorithm contain 10 times more marginals than the decomposable releases resulting from the simulated annealing search described in Section 1.4. Therefore, when the decomposability constraint is dropped, the resulting set of possible releases is much richer. The fact that a non-decomposable frontier is 10 times larger than a decomposable frontier that satisfies the same constraints tells us that decomposability is a very restrictive constraint.

The released marginals for these three thresholds have dimension six or smaller. For thresholds 4 and 5, only one two-way marginal, $[7,8]$, is not released. This marginal is contained in the greedy frontier for threshold 3. From the summaries presented in tables below we learn that, if we were considering the data in this table to be the entire population rather than a sample, almost all the three-way marginals would be releasable.

We can modify the greedy algorithm so that the hierarchical frontier identified

Dimension	Released Marginals	Total Number of Marginals	Percent
1	0	16	0%
2	0	120	0%
3	0	560	0%
4	338	1,820	18.57%
5	1,176	4,368	26.92%
6	55	8,008	0.69%

Frontier obtained from the greedy procedure for $\beta = 4$. The total number of sub-tables in this frontier is $1,569$.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	119	120	99.17%
3	546	560	97.5%
4	1,531	1,820	84.12%
5	1,396	4,368	31.96%
6	55	8,008	0.69%

Breakdown of the released set of sub-tables corresponding to the frontier in Table 1.6. Compare with the released set corresponding with the decomposable frontier $\mathscr{RF}(\beta = 4)$–see Table 1.2. The total number of released sub-tables is $3,663$.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	0	16	0%
2	0	120	0%
3	5	560	0.89%
4	405	1,820	22.25%
5	1,110	4,368	25.41%
6	17	8,008	0.21%

Frontier obtained from the greedy procedure for $\beta = 5$. The total number of sub-tables in this frontier is $1,537$.

Dimension	Released Marginals	Total Number of Marginals	Percent
1	16	16	100%
2	119	120	99.17%
3	545	560	97.32%
4	1,480	1,820	81.32%
5	1,189	4,368	27.22%
6	17	8,008	0.21%

Breakdown of the released set of sub-tables corresponding to the frontier in Table 1.8. Compare with the released set corresponding with the decomposable frontier $\mathscr{RF}(\beta = 5)$–see Table 1.3. The total number of released sub-tables is $3,366$.

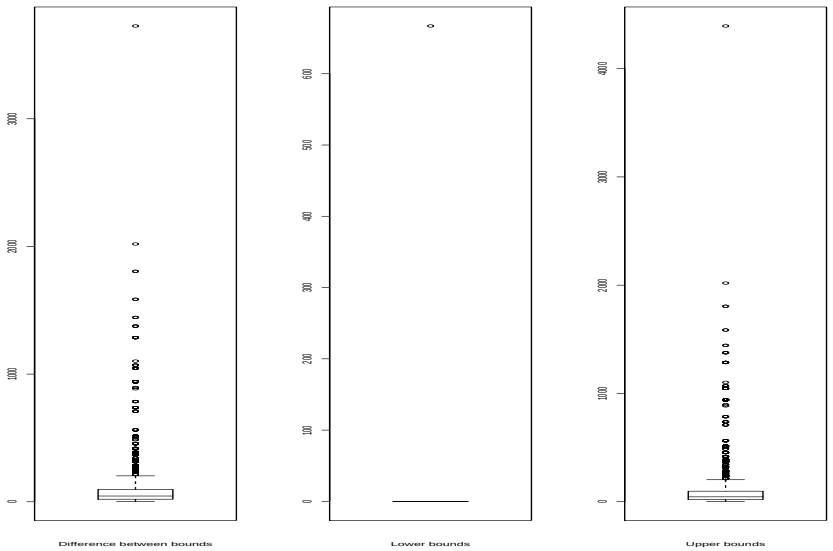

FIGURE 1.3
Boxplots with the bounds for the non-zero cells determined by the frontier $\mathscr{RF}(\beta = 3)$.

includes the MSSs of well fitting log-linear models, provided that these MSSs are simultaneously releasable according to the risk criteria employed. It is sufficient to put the MSSs at the top of the list $\mathscr{L}$, followed by the rest of the marginals in decreasing order of their critical widths. This straightforward approach maximizes the utility of a release from the point of users trying to model the data in the full cross-classification.

1.6 Bounds

In this section, we provide details on the bounds determined by the decomposable and greedy frontiers associated with a threshold equal to "3."

1.6.1 Bounds in the Decomposable Case

We calculated the bounds corresponding to the frontier $\mathscr{RF}(\beta = 3)$ by employing the formulas described in Dobra and Fienberg (2002)–see Figure 1.3.

The upper bounds are strictly bigger than the lower bounds for all the cells in the table. The sum of the differences between the upper and lower bounds for the non-

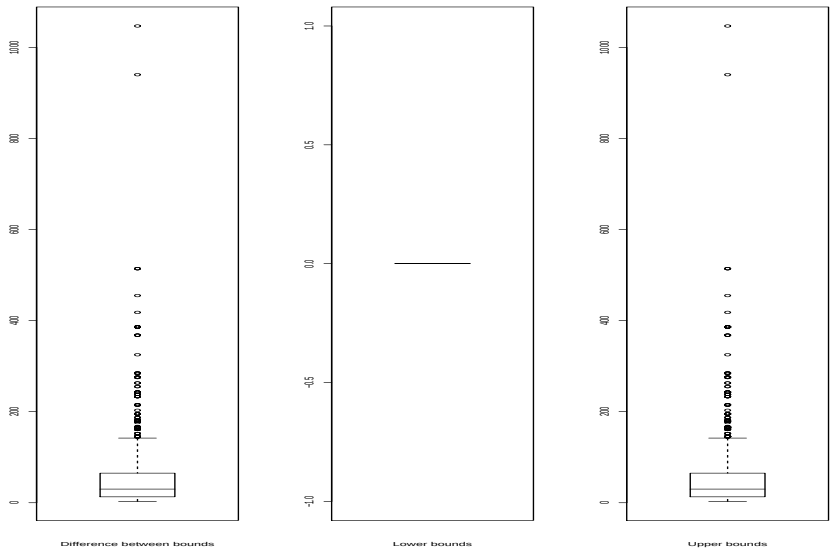

FIGURE 1.4
Boxplots with the bounds for the cells having a count of 1 **determined by the frontier** $\mathscr{RF}(\beta = 3)$**.**

zero cells is $345,534$. All the cells but one have lower bounds equal to 0. The only cells with a non-zero lower bound is the $(0,0,\ldots,0)$ cell in the table, and this lower bound is equal to 667. This cell contains the largest count in the table and consequently has the largest upper bound and the largest difference between the bounds. The minimum value for the upper bounds is 3 and is attained for 11 cells. There are 36 cells with an upper bound of 4, 27 cells with an upper bound of 5 and 55 cells with an upper bound of 6. All the corresponding lower bounds are 0.

In Figure 1.4 and Figure 1.5 we give the bounds associated with the small count cells of 1 and 2, respectively. All the lower bounds for these cells are zero.

1.6.2 Bounds in the Non-decomposable Case

By employing the generalized shuttle algorithm, we calculated the bounds associated with the greedy frontier from Table 1.4–see Figure 1.6.

A number of the $24,148$ cell containing non-zero counts have the upper bounds equal to the lower bounds. However, for the cells having a count of 1, the minimum difference between the bounds is 3 (there are 14 cells for which this minimum is attained), while the minimum difference between the bounds for the cells having a count of 2 is 4 (only two cells have this property).

The sum of the differences between the upper and lower bounds for the non-zero cells is $249,759$, hence the bounds are tighter than the bounds for the decomposable

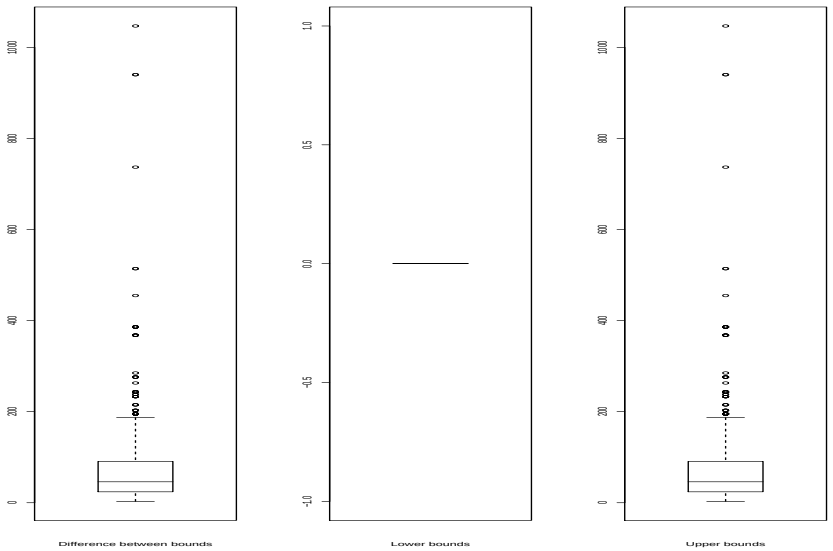

FIGURE 1.5
Boxplots with the bounds for the cells having a count of 2 **determined by the frontier** $\mathscr{RF}(\beta = 3)$**.**

frontier $\mathscr{RF}(\beta = 3)$. Moreover, 22 non-zero cells have lower bounds greater or equal to 1. Again, the first cell in the table has the largest lower bound, the largest upper bound and the largest difference between the bounds.

The lower bounds for the small count cells of 1 or 2 are all zero–see Figure 1.7 and Figure 1.8.

1.7 Discussion

In this paper we presented two methods for determining a releasable frontier. The first method that computes a decomposable frontier is fast and will work for arbitrarily large tables with any number of dimensions and with millions of cells. The scalability of this approach relates to the fact that it is based on computing bounds based on *formulas* whose usage involve little or no computational effort. The only drawback of using this method is that the decomposability can be a serious constraint in many situations: in our example, the size of the frontiers generated by the two methods differed by an order of magnitude.

The second method relaxes this assumption and computes a hierarchical frontier that could have any structure. The first step in applying this method is calculating

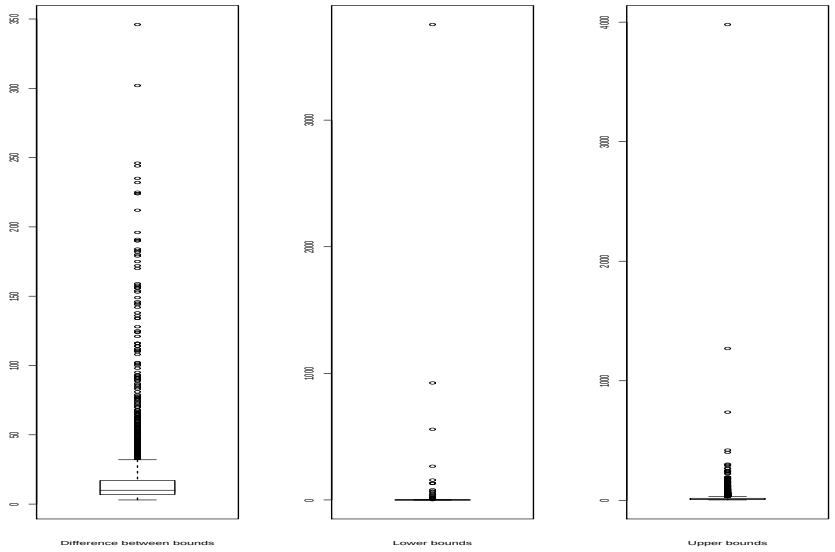

FIGURE 1.6
Boxplots with the bounds for the non-zero cells determined by the frontier from Table 1.4.

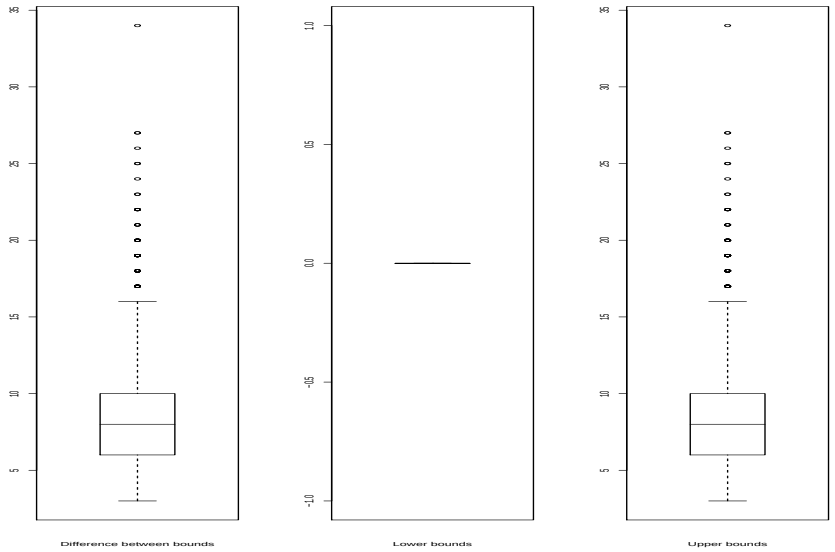

FIGURE 1.7
Boxplots with the bounds for the cells having a count of 1 **determined by the frontier from Table 1.4.**

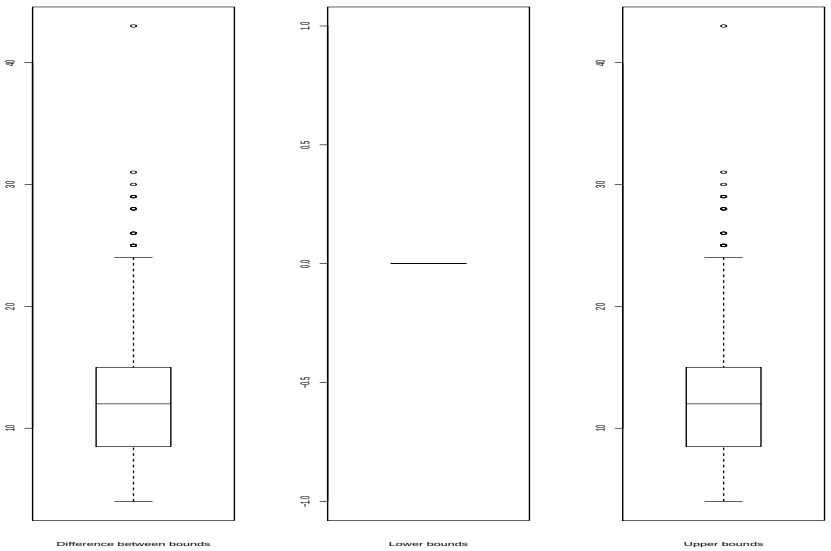

FIGURE 1.8
Boxplots with the bounds for the cells having a count of 2 **determined by the frontier from Table 1.4.**

the critical widths associated with each marginal and this calculation also scales to arbitrary multi-way tables since it is based on the same formulas for computing the bounds. The critical widths have another immediate use: it provides us with a consistent way of ranking variables with respect to how problematic they are for disclosure. As we mentioned before, as the critical width of a marginal gets larger, that marginal tends to tighten the bounds less. Therefore, one would expect that the "less problematic" variables will be contained in marginals with larger critical widths.

We define the *disclosure score* associated with each variable cross-classified in the target table to be the *mean* of the critical widths for all the marginals in which that variable belongs to–see Table 1.10. In our running example, each variable belongs to $32,767$ marginals. We ordered the variables in increasing order with respect to their scores. An increase in the score corresponding with a sequence of variables indicates that the variables in that sequence make the bounds less and less tight, therefore those variables become less and less "problematic." In Table 1.10 we grouped the 16 variables in three groups defined by the disclosure scores: Group 1–very problematic variables; Group 2–problematic variables; Group 3–slightly problematic variables. Group 1 contains two variables, 1 (ADL *eating*) and 7 (IADL *doing heavy house work*), which appear to be significantly more "problematic" than the other variables. It seems important to emphasize that this is not in contradiction with our previous findings which identified $[8,7]$ to be the most problematic combination

Group 1	Variable	1	7						
	Disclosure Score	1.82	1.88						
Group 2	Variable	16	8	11	4	10	9	5	2
	Disclosure Score	2.84	2.91	3.01	3.15	3.17	3.23	3.24	3.26
Group 3	Variable	6	3	12	14	15	13		
	Disclosure Score	3.37	3.39	3.52	3.66	3.74	3.85		

Assessing how problematic every variable in the NLTCS dataset is using disclosure scores.

of two variables for disclosure limitation purposes: combinations of several variables have properties different than the properties of each variable taken by itself. Group 2 contains a combination of ADL and IADL variables. The least problematic variables are four IADL variables and ADLs 3 (*getting around inside*) and 6 (*getting to the bathroom or using a toilet*). Based on disclosure scores, other ADL variables appear to be quite ”problematic”, with the most problematic ADL *eating*. The IADL variables appear to be divided into two groups: with more ”problematic” variables number 7 (*doing heavy house work*), 16 (*telephoning*), 8 (*doing light house work*), and 11 (*grocery shopping*), and less problematic variables number 12 (it getting about outside), 14 (*managing money*), 15 (*taking medicine*), and 13 (*traveling*).

While there is a natural gap in disclosure scores between groups 1 and 2, groups 2 and 3 less separate. In fact, the difference in scores between variables 2 and 6 is smaller than that between variables 3 and 12. If we were to place variables 3 and 6 into group 2, then it would contain all of the ADL variables except for variable 1 and along with some IADL variables. The methods we have applied are most directly appropriate when the table of counts presents population data. It is worth remembering that the data in this example come from a sample survey where the sampling fraction is relatively small, and thus the release of the entire table might well be deemed safe by most disclosure limitation standards.

Finally, address the potential utility of marginal tables released using methodology illustrated in this paper. Users would like to be able to perform statistical analyses on and draw the same inferences from the released marginals as they would were they in the possession of the complete dataset. In principle, it is straightforward to assure the consistency of inferences by making sure that the relevant marginals involved in log-linear models that fit the data well were released. But in the present example, there is no unsaturated log-linear model that fits the data well, and alternative models such as the grade of membership model discussed by Erosheva [17, **?**] seem much more appropriate. These alternative models do a far better job of fitting the very large cells in the table, e.g., the cell corresponding to those with no disabilities, and thus these cells may need to be part of any release, along with a set of marginals. This is an issue we hope to pursue in future research.

References

[1] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. *Discrete Multivariate Analysis: Theory and Practice*. M.I.T. Press, Cambridge, MA, 1975.

[2] Bonferroni, C. E. *Teoria statistica delle classi e calcolo delle probabilitá*. Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze, **8**, 1936.

[3] Buzzigoli, L. and Giusti, A. An Algorithm to Calculate the Lower and Upper Bounds of the Elements of an Array Given its Marginals. In *Statistical Data Protection (SDP'98) Proceedings*, pages 131–147, Eurostat, Luxembourg, 1999.

[4] Chen, G. and Keller-McNulty, S. Estimation of Identification Disclosure Risk in Microdata, *Journal of Official Statistics* **14**: 79–95, 1998.

[5] Cox, L. H. Some Remarks on Research Directions in Statistical Data Protection. In *Statistical Data Protection (SDP'98) Proceedings*, pages 163–176, Eurostat, Luxembourg, 1999.

[6] Deshpande, A., Garofalakis, M. and Jordan, M. Efficient Stepwise Selection in Decomposable Models, In J. Breese and D. Koller (Ed)., *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*, 2001.

[7] Dobra, A. Computing Sharp Integer Bounds for Entries in Contingency Tables Given a Set of Fixed Marginals. Tech. Rep., Department of Statistics, Carnegie Mellon University, 2000.

[8] Dobra, A. Measuring the Disclosure Risk for Multi-way Tables with Fixed Marginals Corresponding to Decomposable Log-linear Models. Tech. Rep., Department of Statistics, Carnegie Mellon University, 2000.

[9] Dobra, A. and Fienberg, S. E. Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs. *Proceedings of the National Academy of Sciences*, **97**: 11885–11892, 2000.

[10] Dobra, A. and Fienberg, S. E. Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals with Applications to Disclosure Limitation. *Statistical Journal of the United Nations ECE*, **18**: 363–371, 2001.

[11] Dobra, A. and Fienberg, S.E. (2002). Bounding Entries in Multi-way Contingency Tables Given a Set of Marginal Totals. *Proceedings of Conference on Foundation of Statistical Inference and Its Applications Jerusalem, Israel*, R. Lerch, (ed.), Springer-Verlag, Heidelberg, to appear, 2002.

[12] Dobra, A., Karr, A. F., Sanil, A. P., and Fienberg, S. E. Software Systems for Tabular Data Releases. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*,, in press (2002).

[13] Doyle, P., Lane, J. Theeuwes, J., and Zayatz, L. (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam, 2001.

[14] Duncan, G. T. and Fienberg, S. E. Obtaining Information While Preserving Privacy: a Markov Perturbation Method for Tabular Data. In *Statistical Data Protection (SDP'98) Proceedings*, pages 351–362, Eurostat, Luxembourg, 1999.

[15] Duncan, G. T., Jabine, T. B., and Wolf, V. A. de (Eds.). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, DC, 1993.

[16] Edwards, D. E. and Havranek, T. A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, **72**: 339–351, 1985

[17] Erosheva, E. Grade of Membership and Latent Structure Models with Application to Disability Survey Data. Department of Statistics, Carnegie Mellon University, Ph.D. Dissertation, 2002.

[18] Erosheva, E. Partial Membership Models with Application to Disability Survey Data. In this volume, 2002.

[19] Erosheva, E. A., Fienberg, S. E. and Junker, B. W. Alternative Statistical Models and Representations for Large Sparse Multi-dimensional Contingency Tables. *Annales de la Faculté des Sciences de l'Université de Toulouse Mathématiques*, **11**, in press, 2002.

[20] Fienberg, S. E. Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics*, **10**, pages 115–132, 1994..

[21] Fienberg, S. E. Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. In *Statistical Data Protection (SDP'98) Proceedings*, pages 115–129, Eurostat, Luxembourg, 1999.

[22] Fienberg, S. E. and Makov, U. E. Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data. *Journal of Official Statistics*, **14**: 485–502, 1998.

[23] F̊ienberg, S. E. and Makov, U. E. Uniqueness and Disclosure Risk: Urn Models and Simulation. *Research in Official Statistics* **4**: 23–40, 2001.

[24] Fienberg, S. E., Makov, U. E., Meyer, M. M., and Steele, R. J. Computing the Exact Distribution for a Multi-way Contingency Table Conditional on its Marginals Totals. In *Data Analysis from Statistical Foundations: Papers in Honor of D. A. S. Fraser*, ed. A. K. Md. E. Saleh, Nova Science Publishing, Huntington, NY, pages 145–177, 2001.

[25] Fienberg, S. E., Makov, U. E., and Steele, R. J. Disclosure Limitation Using Perturbation and Related Methods for Categorical Data. *Journal of Official Statistics*, **14**, pages 485–502, 1998.

[26] Fréchet, M. *Les Probabilitiés, Associées a un Système d'Evénments Compatibles et Dépendants. Premiere Partie.* Hermann & Cie, Paris, 1940.

[27] Hoeffding, W. Scale-invariant Correlation Theory. Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin, **5**(3), pages 181–233, 1940.

[28] Joe, H. *Multivariate Models and Dependence Concepts.* Chapman & Hall, New York, 1997.

[29] Katz, A., Ford, A. B., Moskowitz, R. W., Jackson, B. A., and Jaffe, M. W. Studies of illness in the aged. The index of ADL: A standardized measure of biological and psychosocial function. *Journal of the American Medical Association*, **185**, 914–919, 1963.

[30] Keller-McNulty, S. and Unger, E. A. A Database System Prototype for Remote Access to Information Based on Confidential Data. *Journal of Official Statistics*, **14**: 347–360, 1998.

[31] Lauritzen, S. L. *Graphical Models.* Clarendon Press, Oxford, 1996.

[32] Leimer, H. G. Optimal Decomposition by Clique Separators. *Discrete Mathematics*, **113**: 99–123, 1993.

[33] Manton, K. G., Woodbury, M. A. and Tolley, H. D. *Statistical Applications Using Fuzzy Sets*. Wiley, New York, 1994.

[34] Roehrig, S. F., Padman, R., Duncan, G. T., and Krishnan, R. Disclosure Detection in Multiple Linked Categorical Datafiles: A Unified Network Approach. In *Statistical Data Protection (SDP'98) Proceedings*, pages 149–162, Eurostat, Luxembourg, 1999.

[35] Samuels, S. M. A Bayesian, Species-sampling-inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, **14**: 373–383, 1998.

[36] Skinner, C. J. and Holmes, D. J. Estimating the Re-identification Risk Per Record in Microdata. *Journal of Official Statistics*, **14**: 373–383, 1998.

[37] Willenborg, L. and de Waal, T. *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics, Vol. **111**, Springer-Verlag, New York, 1996.

[38] Willenborg, L. and de Waal, T. *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics, Vol. **155**, Springer-Verlag, New York, 2000.