European Journal of Operational Research 250 (2016) 91–100

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Discrete Optimization

Obtaining cell counts for contingency tables from rounded conditional frequencies

Andrew J. Sage[a], Stephen E. Wright[b,*]

[a] Department of Statistics, Iowa State University, Ames, Iowa, United States
[b] Department of Statistics, Miami University, Oxford, Ohio, United States

ARTICLE INFO

Article history:
Received 21 January 2014
Accepted 7 September 2015
Available online 12 September 2015

Keywords:
OR in government
Integer linear programming
Statistical disclosure control
Tabular data
Fast Fourier transform

ABSTRACT

We present an integer linear programming formulation and solution procedure for determining the tightest bounds on cell counts in a multi-way contingency table, given knowledge of a corresponding derived two-way table of rounded conditional probabilities and the sample size. The problem has application in statistical disclosure limitation, which is concerned with releasing useful data to the public and researchers while also preserving privacy and confidentiality. Previous work on this problem invoked the simplifying assumption that the conditionals were released as fractions in lowest terms, rather than the more realistic and complicated setting of rounded decimal values that is treated here. The proposed procedure finds all possible counts for each cell and runs fast enough to handle moderately sized tables.

1. Introduction

Statistical disclosure control is concerned with privacy guarantees when releasing data that might otherwise identify, or reveal information about, specific individuals or organizations. Releasing data in summarized form greatly reduces disclosure risk, but does not eliminate the risk altogether. In particular, a contingency table of frequency counts potentially reveals information if the table includes any cells having small or zero counts. On the other hand, a table is less useful for statistical inference if it overly aggregates information. This trade-off generally leads to the release of a modified or aggregated version of the table that carries somewhat less information than the original. In deciding which particular form to release, a crucial step is to determine the tightest possible bounds that can be inferred about the individual cell counts in the original table. Such bounds can be used to assess both the disclosure risk and the statistical utility of the released information. The present paper addresses the open question of how to calculate these bounds when the contingency table is released in the form of a two-way table of rounded conditional frequencies.

A *two-way contingency table* is a two-dimensional array of whole numbers, and the row and column sums of such an array are referred to as *marginal* counts. The *conditional row probabilities* are given by dividing each entry in the array by the marginal sum for its row. As an example, the contingency tables shown in Tables 1(a) and 1(b) have the same *sample size* (sum of all entries) and also lead to the same conditional row probabilities, which are shown in Table 1(c). Wright and Smucker (2014) showed that Tables 1(a) and 1(b) are the only contingency tables of sample size 48 having the row conditionals shown in Table 1(c). Because Tables 1(a) and 1(b) have the same second row, knowledge of Table 1(c) and the sample size therefore exposes the actual counts in the second row. The question is whether those counts could be obscured somewhat by converting the exact fractions in Table 1(c) into rounded decimal expressions. In Section 3 we show that two-digit rounding does obscure the counts in this example whereas three-digit rounding does not.

Various approaches are used to limit the disclosure risk of contingency tables, such as choosing a more heavily aggregated summary (e.g., marginal counts only), suppressing some cells altogether, or perturbing cell counts slightly (Hundepool, Domingo-Ferrer, Franconi, Giessing, Nordholt, Spicer, & de Wolf, 2012). These often rely on operations research techniques to limit risk and evaluate risk-utility trade-offs (Almeida & Carvalho, 2005; Castro, 2006; 2011; 2012; Cox, 1995; Cox & Ernst, 1982; Fischetti & Salazar-González, 1999; Hernández-García & Salazar-González, 2014; Kelly, Golden, & Assad, 1990; Kelly, Golden, Assad, & Baker, 1990; Muralidhar & Sarathy, 2006; Salazar-González, 2004; 2005; 2006; 2008). Over the past decade researchers have also explored the possibility of releasing observed conditional probabilities, which retain some statistical utility insofar as odds and ratios of odds are preserved (Slavković, 2010). To date, the work on disclosure risk for tables of conditionals has focused primarily on two-way contingency tables, such as two-way

[*] Corresponding author. Tel.: +15135291837.
E-mail address: wrightse@MiamiOH.edu (S.E. Wright).

http://dx.doi.org/10.1016/j.ejor.2015.09.011
0377-2217/© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.

Table 1
Two contingency tables, (a) and (b), with sample size 48 and row conditionals (c).

(a)			(b)			(c)	
3	4	7	9	12	21	3/7	4/7
5	3	8	5	3	8	5/8	3/8
6	9	15	4	6	10	2/5	3/5
10	8	18	5	4	9	5/9	4/9
24	24	48	23	25	48		

rearrangements of multi-way tables in which some subset of the observed variables are treated as the conditions (i.e., predictors), some are treated as the responses, and perhaps others are omitted altogether (by aggregating over their values) (Fienberg & Slavković, 2005; Slavković, 2004; Smucker & Slavković, 2008; Smucker, Slavković, & Zhu, 2012).

Our problem description and methods are framed in terms of two-way contingency tables, but the same approach can also be used to obtain cell bounds on multi-way contingency tables that have been reshaped as two-way tables by designating some variables as predictors, other variables as responses, and perhaps omitting some variables altogether. We refer the reader to Smucker et al. (2012), Wright and Smucker (2013), Wright and Smucker (2014) for more information on how and why such reshaping might be performed.

Here is a concrete example of a multi-way table using public data that nevertheless provides a nice illustration of how several variables might be grouped into predictor, response, or omitted variables. We consider an 8-way table ($N = 48,842$) from the 1993 U.S. Current Population Survey (CPS), a monthly survey that collects demographic and other data of interest. Table 2 gives information about the qualitative variables measured in this study. Imagine that data such as these comprised, say, all the adults in a given small city. There might be some concern in releasing it in full detail. In particular, small cell counts (including zeros) could potentially be used to identify the salary range of specific individuals. Other variables in this data set might also be considered sensitive information for some people, depending on their personal circumstances. Among the eight variables here, we would almost always consider age, race and sex as natural choices for possible predictors, whereas the other variables (except salary) could be considered either a predictor or response depending on the question at hand. Likewise, any of the variables might conceivably be omitted (implying an aggregation of counts). Moreover, levels within a variable might be combined into a smaller number of levels, such as reducing the three age ranges (for variable X_1 of Table 2) to two age ranges. In any case, the value of row conditionals now becomes clearer: over all people in the study who satisfy a given set of demographic predictors, the row conditionals tell us what proportion have a given (say) education level and salary level. The actual counts could well be of less interest than the proportions to decision-makers and the general public. But privacy considerations and statistical inference would require recovering information about those underlying counts from the stated proportions.

Table 2
CPS variables and their levels.

	Variable	Levels
X_1	Age	< 25, 25–55, > 55
X_2	Employment	Gov't, private, self-employed, other
X_3	Education	$<$ HS, HS, college, bachelor, bachelor+
X_4	Marital status	Married, unmarried
X_5	Race	Non-white, white
X_6	Sex	Female, *Male*
X_7	Hours worked	< 40, 40, > 40
X_8	Salary	< 50, 50+

The mathematical structure of the corresponding cell-bounding problem was recently elucidated for the simpler context in which conditional probabilities are presented as *unrounded* fractions Slavković, Zhu, and Petrović, Wright and Smucker (2013). Under that idealization, it was demonstrated that the (upper or lower) bounding problem for each cell can be reduced to an integer linear knapsack problem. Wright and Smucker (2014) subsequently showed that cell bounds and possible counts for an entire two-way table of unrounded conditionals can be obtained quickly with an algorithm that shares intermediate solution information among large groups of cells in the table. They used that capability to explore disclosure risk for various rearrangements of the data represented in Table 2 under the assumption that a data snooper somehow determined the true unrounded fractions for each conditional probability.

The present article examines the two-way cell-bounding problem in the more complicated and realistic setting of *rounded* conditionals. Several of the works cited above (especially Smucker et al., 2012, Slavković, Zhu, & Petrović) provide light commentary, heuristics or preliminary results on issues relating to cell-bounding from rounded conditionals. So far there have been no substantive attempts to provide a general procedure for identifying tightest cell bounds. Here we formulate the bounding problem for each cell as a pair of integer linear optimization problems and show how these can be decomposed into two types of simpler subproblems. One type of subproblem is solvable in closed form and the other can be addressed by adapting some of the ideas presented in Wright and Smucker (2014). The decomposition provides useful knowledge about the cell-bounding problem structure and identifies all possible cell *counts* rather than merely the cell bounds (which is the most that could be obtained with integer programming methods). We also provide simple examples showing that: (a) rounded conditionals can lead to considerably wider tightest bounds than the corresponding unrounded fractions, and (b) a single-digit change in the rounding precision sometimes means the difference between revealing many cell counts and revealing none.

A full discussion of how to evaluate statistical utility and disclosure risk goes beyond the scope of this paper, but the basic issues are as follows. On the one hand, disclosure risk is minimized by avoiding the revelation of small counts (magnitude 3 or smaller, say) in the contingency table, and this in turn suggests that very narrow cell bounds are undesirable. This is easily understood, so our discussion of the examples in this paper tends to focus on disclosure. On the other hand, statistical utility refers to the ability to perform estimation or hypothesis testing and obtain results only slightly weaker than what one could obtain by applying standard statistical methods to the actual table of counts. Intuitively, such utility is maximized by having a more accurate representation of the underlying table, but the picture is more complicated than that. Salazar-González (2004) (also Salazar-González, 2005) has presented a fine overview of statistical disclosure limitation and its trade-offs from an operations research perspective.

Our method finds bounds on row marginals along with information about individual cell counts, with the latter then implying bounds on the remaining marginals. Performing inference on the basis of such bounds is an area of ongoing research, and a few methods have been proposed for the purpose (see Slavković, Zhu, and Petrović, Dobra, 2012 and references cited therein). Given that the implied bounds on marginals could be quite wide when disclosure risk is avoided, we are led to ask how badly data utility might be compromised. Slavković and Lee (2010) have provided some encouraging results on this last point. They proposed a method for disclosure limitation based on the following idea: round the actual row conditional probabilities to nearby decimal-place approximations, rewrite those approximate conditionals as fractions themselves, and then randomly select a table of counts from the collection of contingency tables for which the new fractions are the conditional probabilities. Releasing