
Statistical theory and methods can make the release of partial information from confidential categorical data useful while limiting disclosure risk.

Making the Release of Confidential Data from Multi-Way Tables Count

Stephen E. Fienberg and Aleksandra B. Slavkovic
Department of Statistics
Carnegie Mellon University

Statistical disclosure limitation (SDL) and confidentiality have often been shrouded with a non-statistical veil and the methodology for protecting confidential data has produced problematic outcomes for research data users. Here we describe one possible statistical approach to SDL for data in the form of multi-dimensional contingency tables that illustrates the following points:

- For categorical data the traditional form of reporting has been marginal tables and conditionals.
- Releasing such partial information is compatible with and useful for statistical methods for log-linear models and directed acyclic graphs.
- Interesting new research problems arise in this area.

Some History

From the early part of the twentieth century, confidentiality has been an important element of the mantra of statistical agencies, and it became embedded in the culture of the U. S. Census Bureau with the protections associated with the 1929 Census Act (now known as Title 13 of the U.S. Code).

But the term confidentiality was always thought about in terms of the protection of individual and establishment data and not the release of data to policy makers, researchers, and the public more generally. Moreover, for many confidentiality represented an “absolute” concept and it was not until the 1970s that there was movement towards making statistical thinking more central to the operational implementation of confidentiality protection. More specifically, the President’s Commission on Federal Statistics, issued in 1971, placed special emphasis on confidentiality and subsequently, when the Office of Management and Budget’s Statistical Policy office created the Federal Committee on Statistical Methodology, one of its first activities was a study of confidentiality and statistical disclosure protection (Working Paper No. 2).

Working paper No. 2 was of special interest in part because it signaled for the first time the importance of the trade-off between access and confidentiality and presented Tore Dalenius’s probabilistic notion of disclosure.

The past 25 years have seen the growth of disclosure limitation as a statistical subdiscipline, and the term itself which was a change from that used in the 1970s recognized that about the only way to prevent the total protection of confidentiality was not to release any data at all.



Figure 1: Poster

Thus an agency or a data collector's goal should be to limit the risk of disclosure of information that might prove to be harmful to respondents who provide the information, while at the same time providing as much data to others as possible for analysis.

[Show 1940s Census Bureau confidentiality poster in Figure 1 and the following definition due to Dalenius in a sidebar]

The reader is asked to keep in mind that the concept of disclosure presented here is a very broad one. It would not be desirable to require that there be a zero risk of disclosure, as defined below, in any release of tabulations or microdata files. Such a requirement would end a large proportion of all releases now being made. This would be too great a price to pay for complete elimination of any risk of disclosure.

...

If the release of the statistics S makes it possible to determine the value [of confidential statistical data] more accurately than is possible without access to S , a disclosure has taken place. [Working Paper No. 2, pp. 7 and 9]

Reporting Tabular Data

For as long as most of us can remember, government agencies and social science and public health researchers have reported on the results of observational and experimental data in tabular form,

Table 1: 2000 U.S. Decennial Census Data on Sex, Age, and Race for a Block in Pittsburgh

Sex Age Race	Male		Female	
	Under 18 years	18 Years and over	Under 18 years	18 Years and over
White	4	31	3	32
Black	0	1	0	0
Asian	1	3	2	3
Two or more races	1	0	2	0

often in the form of marginal cross-classifications of counts or as proportions or percentages adding to 1 for a key explanatory variable. In particular, this has been a standard form of reporting the results of sample surveys, typically in the form of 2-way and 3-way tables. This was largely a matter of convenience, because these kinds of tables are easier to fit on a page in a publication, but also because this allowed researchers to calculate the marginal joint relationships between pairs of variables, or even partial relationships conditional on a third variable.

Table 1 reports data from the U.S. decennial census extracted from the AmericanFactfinder website (<http://factfinder.census.gov>) for a block in Pittsburgh. The data show the population data cross-classified by sex, age, and race. Note that because there is only one person in this block who is a Black male and 18 years of age or older, that person is unique in the population and census data for him can possibly be linked to other data resulting in a disclosure. Similarly, there are two other “population uniques” in this table and two counts of “2” that present possible disclosure problems as well. Many government agencies view counts of “3” as similarly problematic.

While some of the American Factfinder data is reported as both counts and percentages for selected categories (e.g., by gender), other statistical agencies report certain table only in percentage or rate form. Figure 2 is such a table from the Bureau of Labor Statistics reporting data from the Current Population Survey.

For both Table 1 and Figure 2, the actual data gathered on the individuals consists of far more than three variables. Thus one of the questions we need to ask is: What kinds of data are releasable from a higher dimensional table that will not raise confidentiality concerns and problems? The other is: Will the released data be useful for statistical inference purposes?

A Clinical Trial Example

We have cast the discussion thus far in the context of government statistical data, but similar issues of confidentiality and usefulness of data arise in other contexts such as epidemiological studies and clinical trials in public health and medicine. In Table 2, we present data from Koch et al. (1983) on the results of a clinical trial on the effectiveness on an analgesic drug, for patients of two different statuses and from two different centers. Given that the individuals in the clinical trial form a “population”, confidentiality questions will focus on the potential harm associated with the release of information on the four cells with counts of “3” in the table, corresponding to two sets of three individuals in ‘Center 1’ and two sets of three individuals in ‘Center 2.’

Table 2. Volunteer rates by sex, race, Hispanic origin, and selected characteristics, September 2002

Selected characteristics	White			Black			Hispanic		
	Total	Men	Women	Total	Men	Women	Total	Men	Women
Age									
Total, 16 years and over	29.4	25.1	33.4	19.2	16.7	21.1	15.7	12.9	18.4
16 to 19 years	28.6	24.3	33.0	18.8	16.3	21.1	18.1	15.3	20.9
20 to 24 years	19.3	15.7	22.9	13.1	9.9	15.8	9.4	7.6	11.3
25 to 34 years	26.8	20.8	32.7	20.2	15.6	24.0	16.9	12.9	21.0
35 to 44 years	37.1	30.8	43.4	22.4	19.1	25.2	20.6	15.7	25.4
45 to 54 years	33.5	29.4	37.6	20.4	19.3	21.2	16.1	15.1	17.1
55 to 64 years	28.8	26.1	31.4	20.6	19.1	21.7	13.2	12.0	14.2
65 years and over	23.9	22.2	25.2	13.9	14.9	13.3	6.9	6.2	7.4
Employment status among persons aged 16 years and over									
Employed	31.4	27.1	36.6	21.9	18.9	24.6	17.0	14.0	21.1
Unemployed	26.5	21.3	32.6	21.5	18.2	24.6	17.9	12.3	25.5
Not in the labor force	25.6	20.1	28.9	14.1	12.1	15.5	12.6	9.0	14.4
School enrollment status among persons aged 16 to 24 years									
Enrolled in high school	32.3	26.0	39.5	18.2	17.0	19.4	19.6	15.6	23.9
Enrolled in college	28.3	25.2	31.1	23.9	19.7	26.4	19.6	19.4	19.7
Not enrolled in school	16.0	13.0	19.1	10.5	8.4	12.7	8.6	7.2	10.3
Educational attainment among persons aged 25 years and over									
Less than a high school diploma	10.5	9.0	11.8	9.2	8.6	9.6	8.4	5.8	11.0
High school graduate, no college ¹	22.8	18.2	26.8	14.1	12.7	15.4	16.3	13.4	19.2
Less than a bachelor's degree ²	34.5	28.9	39.4	26.1	23.2	28.0	25.2	22.9	27.2
College graduate	46.0	40.9	51.4	36.6	33.4	39.1	31.9	27.2	36.4

¹ Includes high school diploma or equivalent.
² Includes the categories of some college, no degree; and associate's degree.

Note: Data on volunteers relate to persons who performed unpaid volunteer activities for an organization at any point from September 1, 2001, through the survey week in September 2002. Details for the above race and Hispanic-origin groups will not sum to totals because data for the "other races" group are not presented and Hispanics are included in both the white and black population groups.

Figure 2: Published 3-way Table with Rates from 2002 Current Population Survey Supplement. Source: Boraas, S. "Volunteerism in the U.S." *Monthly Labor Review*, August 2003 and taken from the U.S. Bureau of Labor Statistics website.

Here there is a specific analytical question of interest: What is the effect of the treatment on the response, controlling for the other two variables?

What we demonstrate here, is that it is possible to release data from this four-dimensional table that would allow an analyst to make proper inferences about the substantive question of interest without fully disclosing the four cells containing counts of 3.

Margins and Log-linear Models

In the 1960s statistical methodologists created the core theory for log-linear models for the analysis of multi-dimensional contingency tables and the theory turned out to fit rather nicely with the reporting practice. A key theoretical result is that the "minimal sufficient statistics" or "data summaries needed for efficient estimation" associated with a log-linear model corresponded to the highest order terms or interactions in the model, e.g., a two-way margin corresponds to a first-order interaction for the corresponding variables, and a three-way margin corresponds to a 2nd-order interaction. The new ideas on log-linear models made clear, however, that an analyst had to use the information in all of the minimal sufficient statistics simultaneously for estimation purposes and not simply proceed piecemeal by looking at the association margin by margin (e.g., see Bishop, Fienberg, and Holland, 1975; Agresti, 2002). Otherwise one might mistakenly infer dependencies among variables that in effect were explained by other dependencies, or even get a reversal of the "sign" associated with the association, as in the phenomenon known as Simpson's paradox.

Log-linear model theory explained how to do the estimation and how to assess the fit of the

Table 2: Results of Clinical Trial for the Effectiveness of an Analgesic Drug. Source: Koch et al. (1983).

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	3	20	5
1	1	Placebo	11	14	8
1	2	Active	3	14	12
1	2	Placebo	6	13	5
2	1	Active	12	12	0
2	1	Placebo	11	10	0
2	2	Active	3	9	4
2	2	Placebo	6	9	3

models to the data in a multi-way table and, while we need not concern ourselves with the details of the methodology here, every major statistical package either includes specific programs for carrying out the calculations or has a generalized linear model program that can be used for this purpose.

The happy confluence of log-linear model theory and the desire to report marginals means that a statistical agency or the researchers carrying out a clinical trial or epidemiological investigation could possibly share partial information in the form of marginals with users (researchers) and still protect the confidentiality of the data in a multi-way table.

To estimate associations the user needs the margins to go with a “good” log-linear model that fits the data well. To check on the model fit we need more data than the minimal sufficient statistics for the model itself, i.e., more margins or at least some of higher-dimension. This more elaborate data release then corresponds to a more complex log-linear model and we can then compare the expected values under the simpler model with the more complex one.

For the data in Table 2, we need to include the margin for the three explanatory variables, i.e., Center by Status by Treatment—we use the notation [CST] as a shorthand for this three-way margin. And virtually all model search procedures would narrow the focus to two models:

1. [CST] [CSR],
2. [CST][CSR][TR].

both of which fit the data well. Model 1 is a special case of model 2 and the likelihood ratio test for the difference between them takes the value $\Delta G^2 = 5.4$ with 2 degrees of freedom, a value that is not significant at the 0.10 level when compared with a chi-squared distribution. Thus one might reasonably conclude that the effect of the treatment on the response is explained through the interactive effect of Center and Status. A key point for the present purposes is that we need three sets of marginal totals to make this inference: [CST], [CSR], and [TR].

Table 3: Upper and Lower Bounds For Cell Entries in Table 2 Given the [CST] and [R] Margins.

Center	Status	Response	Poor	Moderate	Excellent
		Treatment			
1	1	Active	[0,28]	[0,28]	[0,28]
1	1	Placebo	[0,33]	[0,33]	[0,33]
1	2	Active	[0,29]	[0,29]	[0,29]
1	2	Placebo	[0,24]	[0,24]	[0,24]
2	1	Active	[0,24]	[0,24]	[0,24]
2	1	Placebo	[0,21]	[0,21]	[0,21]
2	2	Active	[0,16]	[0,16]	[0,16]
2	2	Placebo	[0,18]	[0,18]	[0,18]

Bounds On Tables Entries Given Marginals

Earlier we noted that the risk of identity disclosure in a table of counts is usually associated with small cell values. Small counts such as “1”, “2” and potentially “3” allow an intruder to match characteristics in table with other databases and learn confidential information. But if we only report selected margins from a multi-way table with such small values, can that information be used to infer values in the cells of the full table?

For two-way tables, statisticians and others have long known how to place bounds on the entries of the table given the (one-way) margins. For an $I \times J$ table with table entries n_{ij} , row margins n_{i+} and column margins n_{+j} , these bounds have the following form:

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}. \quad (1)$$

Thus, if we treat the data in Table 2 as if they come from an 8×3 table and apply equation (1), we get the bounds in Table 3. There are 6,718,227,637,086,252 tables with the same sets of marginal totals and across all of them these are the maximum and minimum values for each of the cell counts. We note that all of the lower bounds in this example are 0 even though this need not be the case in general. Since the uppers bounds are far from the lower bounds and since these bounds correspond to an extremely large collection of tables, an intruder cannot use them to make strong inferences about potentially small cell entries.

Of course the rows of Table 2 correspond to three variables and thus we have computed the bounds for a four-way table given the margins [CST] and [R]. Over the past decade, the ideas on bounds have been extended to multi-way tables given two or more, possibly overlapping margins, and not surprisingly these extensions are linked to the theory of log-linear models. Many special cases have explicit formulas like those in equation (1). For other cases various numerical procedures can produce bounds example by example. Dobra and Fienberg (2000, 2003) give many of the details.

If a cell count is small and the uper bounds is close to the lower bound, the intruder knows with some degree of certainty that there is only a small number of individuals possessing the characteristics corresponding to the cell and this may pose a risk of disclosure of the identity of these individuals.

Table 4: Upper and Lower Bounds For Entries in Table 2 Given the [CST] , [CSR], and [TR] Margins.

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	[0,14]	[1,28]	[0,13]
1	1	Placebo	[0,14]	[6,33]	[0,13]
1	2	Active	[0,9]	[3,27]	[1,17]
1	2	Placebo	[0,9]	[0,24]	[0,16]
2	1	Active	[2,21]	[3,22]	[0,0]
2	1	Placebo	[2,21]	[0,19]	[0,0]
2	2	Active	[0,9]	[0,16]	[0,7]
2	2	Placebo	[0,9]	[2,18]	[0,7]

For the data in Table 2, we observed earlier that the four cell entries of “3” pose potential disclosure risk and we would like to protect them by releasing only subsets of the data in the form of marginal totals. We have explored the possible bounds associated with the release of the [CST] margin and all other possible sets of margins. Table 4 contains the bounds for the sets of margins needed to fit and compare the two log-linear models of analytical interest, [CST][CSR] and [CST][CSR][TR] and now we clearly see several cells with positive lower bounds. As before, all of the upper bounds are reasonably far from the lower bounds except for the (2,1,2,3) cell where the upper and lower bounds are now 0, and perhaps the (2,2,2,3) cell where the bounds are [0,7] in both tables. If we released the [CST], [CSR], and [TR] margins an intruder would be far from certain what entries belonged in the 4 cells containing the value 3.

While it is true that releasing the [CST], [CSR], and [TR] margins allows someone else to carry out the likelihood ratio test to assess the effect of the treatment on the response, releasing even more information would be desirable. There are two additional three-way margins to consider: [CTR] and [STR]. If we also release [STR], then we get the bounds in Table 5, which show that the (1,1,1,3) cell which contains a count of 5 and the (1,1,2,3) cell which contains a count of 8 are identified with certainty. If instead we add the three-way margin [CTR], from Table 6 we see that the count of 4 in the (2,2,1,3) cell and the count of 3 in the (2,2,2,3) are revealed with certainty. So it may be possible to release a bit more information in the form of the the [STR] margin but the release of [CTR] is problematic.

The moral of this example is that: when we are faced with a relatively sparse multi-way contingency table containing small counts that might disclose sensitive information about individuals with reasonably high probability, we still are able to release enough of the marginal totals from the table to allow a statistician to explore relevant questions of inference. This is like protecting our statistical cake from disclosure while still allowing others to eat enough of it to enjoy the party!

But What About Conditionals?

This idea of partial releases in the form of sets of margins can be extended to other types of data summaries such as marginal tables of rates, that is conditional or relative observed frequencies

Table 5: Upper and Lower Bounds For Entries in Table 2 Given the [CST], [CSR], and [STR] Margins.

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	[0,13]	[10,23]	[5,5]
1	1	Placebo	[1,14]	[11,24]	[8,8]
1	2	Active	[0,6]	[7,20]	[9,16]
1	2	Placebo	[3,9]	[7,20]	[1,8]
2	1	Active	[2,15]	[9,22]	[0,0]
2	1	Placebo	[8,21]	[0,13]	[0,0]
2	2	Active	[0,6]	[3,16]	[0,7]
2	2	Placebo	[3,9]	[2,15]	[0,7]

Table 6: Upper and Lower Bounds For Entries in Table 2 Given the [CST], [CSR], and [CTR] Margins.

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	[0,6]	[9,28]	[0,13]
1	1	Placebo	[8,14]	[6,25]	[0,13]
1	2	Active	[0,6]	[6,25]	[4,17]
1	2	Placebo	[3,9]	[2,21]	[0,13]
2	1	Active	[6,15]	[9,18]	[0,0]
2	1	Placebo	[8,17]	[4,13]	[0,0]
2	2	Active	[0,9]	[3,12]	[4,4]
2	2	Placebo	[0,9]	[6,15]	[3,3]

for a margin (for example, see Figure 2). Until recently nothing was known, however, about the effect of their release on confidentiality. Furthermore, releasing of conditional distributions for higher-dimensional contingency tables could be useful for researchers interested in assessing causal inference using directed acyclic graphs while still maintaining confidentiality. We are currently exploring the theory associated with such releases (see Slavkovic, 2004, and Slavkovic and Fienberg, 2004) and illustrate a few of the ideas here.

If we go back to our example, clearly we can explore the question of treatment effect by using the full conditional distribution of R given C, S, and T—we use the notation $[R|CST]$ to represent this information. If we also have the margin [CST], we can clearly reconstruct the full four-way table! But if we only had the conditional and the size of the experiment we could still put bounds on cell entries and for most of the cells we can actually deduce the exact cell counts.

Next suppose that [CSR] and [TR] are available and that the researchers also release $[T|CS]$ believing that the relative frequencies offer more protection than the three-way marginal [CST]. It is easy to see that this is equivalent to publishing the [CST], [CSR] and [TR] margins; from [CSR] we can get the [CS] margin which together with $[T|CS]$ gives the [CST] margin. We also

get the same bounds by publishing [CS|T] along with [CSR] and [TR]! What is happening in this example is that the release of the margin [CSR] allows for the reconstruction of other margins from their corresponding conditionals. For higher-way tables we don't have the same types of constraints and conditionals can become a useful tool for releasing more data than one might otherwise have considered based on marginals alone (Slavkovic 2004).

Conclusions and Where to Learn More

We have tried to make the argument in this article that statistical disclosure limitation methodology is inherently statistical and that it is essential to understand the ways in which others would like to use released data for analysis purposes. We have focused on the special case of data in the form of counts in a multi-way contingency table and the disclosure limitation method of releasing partial information in the form of marginal tables. These turn out to be essential elements needed by analysts working with log-linear or logit models. The happy confluence of a stream of research on disclosure methods and another stream of research on log-linear model theory into a river that joins confidentiality with analysis illustrates our central argument.

In their article in this issue of *Chance*, Duncan and Stokes discuss the notion of the trade-off between risk and utility. One way to view the example here is as an informal illustration of the RU-confidentiality map. In our example, one gets essentially the maximal amount of information in the data while also providing adequate confidentiality protection.

We also introduced the notion of other possible elements that might be part of a data release in addition to margins. Our discussion of conditional tables was intended to whet the appetite of statisticians who might be interested in working in this emerging area of research. Moving the ideas from tables of counts and margins to other types of statistical data and data summaries remains a challenge.

For further background on log-linear models and methods for their analysis we refer the reader to Bishop, Fienberg, and Holland (1975) or Agresti (2002). Our example comes from Koch et al. (1983). The results on bounds used in the analysis of the example derive from the methods described in Dobra and Fienberg (2000, 2003) and Dobra, Fienberg, and Trottini (2003), and the extension to conditionals is work in progress, some of which can be found in Slavkovic (2004) and Slavkovic and Fienberg (2004).

References

- [1] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA.
- [2] 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(22): 11885–11892, 2000.
- [3] Dobra, A. and Fienberg, S.E. Bounding entries in multi-way contingency tables given a set of marginal totals. In Y. Haitovsky, H.R. Lerche, and Y. Ritov, editors, *Foundations of Statistical*

Inference: Proceedings of the Shores Conference 2000, pages 3–16. Springer-Verlag, Berlin, 2003.

- [4] Dobra, A., Fienberg, S.E. and Trottini, M. Assessing the risk of disclosure of confidential categorical data. In J. Bernardo et al., editors, *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics*, pages 125–14. Oxford University Press, Oxford, 2003.
- [5] Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22. Subcommittee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington, DC. <http://ntl.bts.gov/docs/wp22.html>.
- [6] Koch, G.G., Amara, J., Atkinson, S. and Stanish, W. (1983). Overview of categorical analysis methods. *SAS-SUGI*, 8: 785–795.
- [7] Slavkovic, A.B. *Statistical Disclosure Limitation Beyond the Margins*. Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2004.
- [8] Slavkovic, A.B. and Fienberg, S.E. (2004). Bounds for cell entries in two-way tables given conditional frequencies. *To appear in Privacy in Statistical Databases '2004*, Josep Domingo-Ferrer and Viceng Torra, eds. Lecture Notes in Computer Science, Springer-Verlag, New York.