A Data Disclosure Policy for Count Data Based on the COM-Poisson Distribution

Joseph B. Kadane

Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, kadane@stat.cmu.edu

Ramayya Krishnan

The Heinz School of Public Policy and Management, Carnegie Mellon University,
Pittsburgh, Pennsylvania 15213, rk2x@andrew.cmu.edu

Galit Shmueli

Department of Decision and Information Technologies, Smith School of Business, University of Maryland,
College Park, Maryland 20742, gshmueli@rhsmith.umd.edu

Count data arise in various organizational settings. When the release of such data is sensitive, organizations need information-disclosure policies that protect data confidentiality while still providing data access. In contrast to extant disclosure policies, we describe a new policy for count tables that is based on disclosing only the sufficient statistics of a flexible discrete distribution. This distribution, the COM-Poisson, well approximates Poisson counts but also under- and over-dispersed counts. The sufficient statistics mask the exact cell counts and often also the table size. Under the scenario of a data holding agency and a data snooper, we show that this policy has low disclosure risk with no loss of data utility: Usually, many count tables correspond to the disclosed sufficient statistics. Furthermore, these count tables are equally likely to be the undisclosed table. Finding these solutions requires solving a system of linear equations, which are underdetermined for tables with more than three cells, and can be computationally prohibitive for even small tables. We also consider cell-specific interval bounds, a commonly used disclosure limitation policy, and compare them to our policy. We describe several types of snooper knowledge, their integration with the disclosed statistics, and implications. Applying this policy to three real data sets, we illustrate the low associated disclosure risk.

Key words: sufficient statistics; Conway-Maxwell-Poisson; disclosure risk; data snooper
History: Accepted by Wallace J. Hopp, information systems; received April 17, 2005. This paper was with the authors $3\frac{1}{2}$ months for 3 revisions.

1. Introduction

The problem of data disclosure from statistical databases has become central in recent years. There are great efforts invested in promoting data access and publishing statistics while protecting privacy. The problem is how a data holder can disclose information that is sufficient for decision making and statistical analysis without disclosing the actual data. Disclosure of data can compromise privacy, confidentiality, and national interests (Sweeney 2002a).

The range of examples is wide: health-care management where medical information from multiple sources is combined and disclosure of the combined data can compromise patients, doctors, and HMOs (Chowdhury et al. 1999), data released by the census bureau and other federal agencies for purposes such as policy making, and data released by financial and other private institutions to researchers for the purpose of fraud detection. The problem has become especially acute due to the advances in data storage and linkage through relational databases. Garfinkel

et al. (2002) mention that according to several surveys, privacy compromise is what Americans fear most in the new millennium. The scenario we deal with in this paper consists of a data holding agency that wants to release data to legitimate users in way that maximizes data utility while protecting confidentiality, and on the other side, a data snooper who wants to uncover the confidential data. The disclosure risk occurs when the released information allows the data snooper to infer part or all of the confidential data. This has been termed "inferential disclosure," to differentiate it from direct disclosure that results from unauthorized access such as password breaking (Chowdhury et al. 1999). With today's computational power, data snoopers use sophisticated models to infer confidential attributes from the released information (Sarathy and Muralidhar 2002).

In many cases, de-identifying respondents in a data set is not sufficient to protect privacy because individuals can be re-identified by linking or matching the data to other data or by looking at unique

characteristics found in the released data (Sweeney 2002a). Common disclosure limitation techniques that are aimed at reducing the disclosure risk are based on altering or reducing the data to be released (Dobra et al. 2003, Sweeney 2002b, Domingo-Ferrer and Torra 2005). Most disclosure limitation techniques are based on one of the following principles (Domingo-Ferrer et al. 2002, Duncan et al. 2001, Fienberg et al. 1998, Gonzalez and Cox 2005): cell perturbation, cell suppression, rounding, table redesign, sampling, swapping, and simulation.

In this paper, we suggest a new disclosure policy for one-dimensional count tables. The policy is based on disclosing only the sufficient statistics of a flexible discrete distribution that can well approximate Poisson counts but also under- and overdispersed counts. Overdispersion is prevalent in applications where there is "contagion," or a chain reaction. One example is the number of infections resulting from hospitalization within a hospital. The public health system might require hospitals to report such information to detect large-scale epidemics in the region. However, a hospital might be reluctant to report the exact count table because of class-action lawsuits that might arise from uncovering infections that involved a large number of patients. Another example is the number of injuries per accident by cars that belong to a large taxi service. The Bureau of Transportation Statistics might want to collect the information on injury distribution among taxi riders to change policies. However, taxi companies might be reluctant to report accidents with large numbers of injuries for fear of loss of reputation.

Our proposed disclosure policy masks the sensitive count table well in the sense that there are multiple possible count tables that correspond to the disclosed statistics, even for small count tables. On the other hand, the sufficient statistics contain all the necessary information required for statistical estimation and decision making in the contexts described above.

The most common discrete distribution used for data approximation is the Poisson distribution. If data follow a Poisson distribution, it is sufficient to disclose only the number of observations n and their sum S_1. Given these statistics, to reconstruct the original data, we use the fact that the n integers must sum up to S_1. It turns out that the number of solutions is given by $\binom{S_1+n-1}{n-1}$ (see Feller 1968, p. 38). Burridge (2003) suggests an "information preserving statistical obfuscation" method for releasing data that is based on computing the sufficient statistics and then generating simulated data from the conditional distribution of the data given the sufficient statistics. For discrete data, he suggests a Poisson model. The main drawback of this method is that they assume a Poisson fit, which is limiting (for instance, it is not suitable

for over- and under-dispersed data). In the following, we introduce the Conway-Maxwell-Poisson (COM-Poisson) distribution, a 2-parameter generalization of the Poisson distribution[1] (Shmueli et al. 2005). The COM-Poisson distribution has an additional sufficient statistic, $S_2 = $ the sum of $\log(x_i)!$. From a data utility perspective, these sufficient statistics provide legitimate users interested in statistical inferences with the same data utility as would access to the raw data. When this additional information is disclosed, the number of solutions will always be less than in the Poisson case. However, it does not compromise the raw data in the sense that the probability of reconstructing the original count table is generally low. Furthermore, even if the goal of the snooper is to find bounds for different counts, these bounds tend to be wide. To measure the associated disclosure risk, we use the cell-level entropy-based measure by Domingo-Ferrer et al. (2002) and also develop a global entropy-based measure that is a function of the number of solutions.

We start by describing the COM-Poisson distribution in §2. Next, we describe the reverse relationship leading from the disclosed sufficient statistics to the undisclosed sensitive count data in §3. This section introduces the snooper's decision problem in this setting, and we derive the number of solutions and their distribution. We discuss univariate interval bounds on cells and their limitation, and describe disclosure-risk measures based on these results. Section 4 illustrates the policy applied to three real data sets, showing its advantages for different sizes of count tables. The paper concludes with managerial implications and future directions in §5.

2. The COM-Poisson Distribution

The COM-Poisson probability function is given by

$$P(X = x) = \frac{\lambda^x}{(x!)^\nu} \cdot \frac{1}{Z(\lambda, \nu)}, \quad x = 0, 1, 2, \ldots, \quad (1)$$

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \quad (2)$$

for $\lambda > 0$ and $\nu \geq 0$. This formulation allows for a nonlinear decrease in ratios of successive probabilities in the form

$$\frac{P(X = x - 1)}{P(X = x)} = \frac{x^\nu}{\lambda} \quad (3)$$

and is thus a generalization of the Poisson distribution, as well as a bridge between the geometric ($\nu = 0$)

---

[1] The basic approach we propose can be applied in conjunction with other count data distributions such as the negative binomial distribution. We have chosen to demonstrate our ideas with the COM-Poisson distribution because it is a flexible distribution that is suitable both for over- and under-dispersed data.

and Bernoulli ($\nu \to \infty$) distributions. The ν parameter determines the level of under- or over-dispersion relative to the Poisson distribution (where $\nu = 1$).

The likelihood for a set of n independent and identically distributed (i.i.d.) observations $x_1, x_2, \ldots, x_n$ is

$$L(x_1, x_2, \ldots, x_n \mid \lambda, \nu) = \frac{\prod_{i=1}^{n} \lambda^{x_i}}{(\prod_{i=1}^{n} x_i!)^{\nu}} Z^{-n}(\lambda, \nu)$$

$$= \lambda^{\sum_{i=1}^{n} x_i} e^{-\nu \sum_{i=1}^{n} \log(x_i!)} Z^{-n}(\lambda, \nu)$$

$$= \lambda^{S_1} e^{-\nu S_2} Z^{-n}(\lambda, \nu), \quad (4)$$

where $S_1 = \sum_{i=1}^{n} x_i$ and $S_2 = \sum_{i=1}^{n} \log(x_i!)$. By the factorization theorem, (S_1, S_2) are sufficient statistics for $x_1, x_2, \ldots, x_n$. Furthermore, (4) displays the COM-Poisson distribution as a member of the exponential family. For parameter estimation, see Shmueli et al. (2005) and Kadane et al. (2006).

3. Reconstructing the Count Table from the Sufficient Statistics

3.1. The Snooper's Decision Problem

Denote by f_j the frequency of value j in the data, $j = 0, 1, 2, \ldots, J$. Let us assume that the goal of the data holder is to disclose only as much information about the distribution of the counts that is needed for estimation without disclosing the exact f_j values. Because sufficient statistics contain the minimal information needed for estimating a distribution's parameters, the question is: Given the sufficient statistics S_1, S_2, and the sample size n, to what extent can the original counts f_j, $j = 0, 1, \ldots, J$ be reconstructed? Note that we assume that J is unknown. Thus, a snooper who is trying to reconstruct the count table from the sufficient statistics must solve a system of equations. This requires an estimate or bound on J. In addition, given that there are usually multiple solutions to these equations, it is important to know their distribution. We address these points next.

3.2. Distribution of Solutions

An important issue is the distribution of the integer solutions. This influences the informativeness of the solution space. If there are many solutions but only a few have a high probability of being the correct ones, then the disclosure risk is much larger than if the solutions were equally likely. It turns out that for the COM-Poisson distribution, as is true for all families of distributions comprising an exponential family, the following theorem holds:

THEOREM 1. *The distribution of the data given the sufficient statistics is uniform among all possible integer tables agreeing with the sufficient statistics.*

PROOF. Let $S = (S_1, S_2) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} \log(X_i!))$ be the sufficient statistics and $f = (f_0, f_1, \ldots, f_J)$ be the counts of the data values $0, 1, \ldots, J$. By definition of sufficiency, $P(f = x \mid S = s, \lambda, \nu)$ does not depend on λ and ν. Furthermore,

$$P(f = x \mid S = s, \lambda, \nu) = \frac{P(f = x, S = s \mid \lambda, \nu)}{P(S = s \mid \lambda, \nu)}$$

$$= \frac{P(f = x \mid \lambda, \nu)}{P(S = s \mid \lambda, \nu)}$$

$$= \frac{\lambda^{S_1} e^{-\nu S_2} Z^{-n}(\lambda, \nu)}{\sum_{\tilde{f}} \lambda^{S_1} e^{-\nu S_2} Z^{-n}(\lambda, \nu)}, \quad (5)$$

where the sum is over all $\tilde{f}$ compatible with (s_1, s_2). Let the number of such $\tilde{f}$ be m_s. Then,

$$P(f = x \mid S = s, \lambda, \nu) = \frac{1}{m_s} \quad (6)$$

and is uniformly distributed. □

This means that regardless of the number of solutions, there is no information indicating that some solutions are more likely than others.

3.3. Bounding J

When $J \leq 1$ (i.e., all observations are exclusively zeros or else they are binary), the actual count table can be immediately reconstructed from the sufficient statistics. In this case, $S_2 = 0$ and S_1 gives the count of ones (the number of zeros is $n - S_1$). For $J > 1$, we use prime factorization. We show that J can be bounded and that the number of linear equations is a function of the number of prime numbers below this bound.

The information contained in S_2 is useful for bounding the number of cells J. Consider the smallest prime number p' such that p' does not occur among the prime factors of e^{S_2}.[2] Then, $J < p'$ because the factorial of every number greater or equal to p' must include p' among its factors. Although e^{s_2} is usually very large, these numbers have the special feature that they consist of many low primes, and therefore finding p' is not hard. When $J \leq 3$, we get the exact relation $J = p' - 1$ because the numbers 1, 2, 3 are all prime numbers. This means that disclosing the sufficient statistics when $J \leq 3$ also discloses J itself. When $J = 2$, we can use the three equations relating (n, S_1, S_2) with (f_0, f_1, f_2) to completely uncover the actual counts.

3.4. Solution Space

In this section, we show that using the fact that every integer has a unique prime factorization, the dimension of solutions depends only on the difference between p' and J.

---

[2] Note that S_2 is equal to the natural log of the products of the factorials of the observations, which are integers. Therefore, e^{S_2} is an integer.

Let $n(p)$ be the exponent of p in the prime factorization of e^{S_2} for each $p \in \mathscr{P}(p')$, the set of prime numbers less than p'. By construction, $n(p) \geq 1$ for each such p. We can then express e^{S_2} as

$$e^{S_2} = \prod_{k=2}^{p'-1}(k!)^{f_k} = \prod_{p \in \mathscr{P}(p')} p^{n(p)}. \qquad (7)$$

Now

$$\prod_{k=2}^{p'-1}(k!)^{f_k} = \prod_{k=2}^{p'-1}(k)^{\sum_{j=k}^{p'-1} f_j} = \prod_{k=2}^{p'-1}(k)^{g_k}, \qquad (8)$$

where $g_k = \sum_{j=k}^{p'-1} f_j$. For each k, $2 \leq k \leq p'-1$, let prime p have exponent $a(k, p)$ in the prime factorization of k, so that

$$k = \prod_{p \in \mathscr{P}(p')} p^{a(k, p)}. \qquad (9)$$

Note that $a(k, k) = 1$ and $a(k, p) = 0$ for $k < p$.

Substituting,

$$e^{S_2} = \prod_{p \in \mathscr{P}(p')} p^{n(p)} = \prod_{k=2}^{p'-1}(k)^{g_k} = \prod_{k=2}^{p'-1} \prod_{p \in \mathscr{P}(p')} p^{a(k, p)}$$

$$= \prod_{p \in \mathscr{P}(p')} p^{\sum_{k=2}^{p'-1} g(k)}. \qquad (10)$$

Equating exponents of primes, we have the fundamental equations

$$n(p) = \sum_{k=2}^{p'-1} a(k, p) g_k = \sum_{k=2}^{p'-1} a(k, p) \sum_{j=k}^{p'-1} f_j. \qquad (11)$$

The summation is over the set $2 \leq k \leq j \leq p'-1$, so

$$n(p) = \sum_{j=2}^{p'-1} f_j \sum_{k=2}^{j} a(k, p) \quad \text{for } p \in \mathscr{P}(p'). \qquad (12)$$

In other words, the coefficients f_k in $n(p) = \sum_{k=1}^{p'} \alpha_k f_k$ are the number of prime factors of p that would be contributed by $k!$.

In addition to the equations in (12), there are two additional constraints on the frequencies:

$$
\begin{aligned}
n &= f_0 + f_1 + f_2 + \cdots + f_J, \\
s_1 &= f_1 + 2f_2 + \cdots + Jf_J.
\end{aligned}
\qquad (13)
$$

In practice, this means that we must factor very large numbers (e^{s_2} is usually very large). As mentioned earlier, these numbers have the special feature that they consist of many low primes. Therefore, a simple program that first divides the number repeatedly by two, then by three, etc. will yield the required factorization in no time. However, there are two major issues: If we disclose the value of e^{s_2}, which can be very large even for moderate count tables, it might be impossible to represent it in less than 64 bits, the limit that most computer programs can handle. A solution is to disclose e^{s_2} in the form of its factorization, i.e., $2^{n(2)} \cdot 3^{n(3)} \cdots (p')^{n(p')}$. Given that the

COM-Poisson's structure and how its sufficient statistics are computed are known to the user, this form does not disclose any more information than the exact value of e^{S_2}.

An alternative is to disclose S_2 with a limited number of decimals. Denote the truncated value by $\tilde{S}_2$. Recall that S_2 is a sum of log values, and is therefore a real number with many decimal digits. Even if the number of disclosed decimal digits d is large, the truncation will have a large impact on e^{S_2}, such that $|e^{S_2} - e^{\tilde{S}_2}|$ can be very large. This introduces more uncertainty into the count table reconstruction. We can bound the real value of e^{S_2} by the interval $e^{\tilde{S}_2} \pm 10^{-d}$. Within this interval, there are only a few integers that are factorizations of all primes until p'. However, p' is unknown! One definite case is if $S_2 > 0$, where we can immediately rule out all odd numbers as possible values. The uncertainty in the value of e^{S_2} leads to multiple sets of solutions, one for each legitimate factorization. The question is then, given the combined set of solutions from the multiple sets, what is the distribution of solutions? There are two choices: the first is to treat all solutions equally disregarding which set they came from. Alternatively, we can treat each set of solutions as equally likely and then weight each solution according to the ratio between the number of solutions in its set and the total number of solutions.

Note that disclosing S_2 in this manner does not have a similar effect on decision making using the sufficient statistics. In fact, the disclosure policy would like to choose the format used to disclose S_2 to be one which adds uncertainty to the reconstruction of count tables while not affecting decision making significantly.

Regarding the solution space, the number of equations is determined by the number of prime numbers below p', and the number of unknown frequencies is p'. Table 1 lists the number of equations and unknown frequencies for $J < 17$. This gives a sense of the magnitude of under-determination of the problem.

Another important point is that only the left-hand side of the equations in (12) and (13) depends on the actual data, whereas the right-hand side depends only on p'. In other words, $a(k, p)$ are independent of the data and can be computed in general. Table 2 gives the values of these coefficients for $k, p \leq 13$.

Table 1 Number of Equations and Unknown Frequencies as a Function of J

J	1	2	3–4	5–6	7–10	11–12	13–16
No. of equations	2	3	4	5	6	7	9
No. of frequencies (p')	2	3	5	7	11	13	17

Table 2   Values of $a(k, p)$ for $k, p \leq 13$

					k						
p	2	3	4	5	6	7	8	9	10	11	12	13
2	1		2		1		3		1		2	
3		1			1			2			1	
5				1					1			
7						1						
11										1		
13												1

3.5. Computing Interval Bounds

When the number of cells J is large, the number of solutions can be very large and it is therefore computationally challenging to find all solutions. In such cases, an alternative is to compute bounds on each cell. In some cases, the objective of the snooper is exactly to find such bounds (rather than the solutions). There has been much work on the interval inference problem (see, for example, Chowdhury et al. 1999 and Li et al. 2002). This is usually much easier computationally than computing all the solutions because it involves an optimization task with linear and nonnegative integer constraints. However, univariate interval bounds are computed separately for each cell, independently from other cells, when in truth the counts in the different cells are dependent. This is similar to examining the marginal distributions of a multivariate distribution, thereby ignoring the multivariate structure. In practice, this means that the bounds might suggest solutions that do not exist. For example, consider a case where one of the equations is $f_5 + f_6 = 1$. The LP solution would yield the bounds $[0, 1]$ for each of f_5 and f_6, whereas the pair $f_5 = 1, f_6 = 1$ is not a solution.

To assess the risk associated with univariate cell bounds, Domingo-Ferrer et al. (2002) proposed an entropy-based disclosure-risk (DR) measure. The DR for cell i, as formulated in Boyens et al. (2004), is given by

$$\text{DR}_i = 1/\log_2(U_i - L_i). \tag{14}$$

This is the reciprocal of the conditional entropy based on the assumption that the distribution of values within the interval $[L_i, U_i]$ is uniform. This measure is constructed separately for each cell i, without relation to other cells. It also assumes a uniform distribution of values within the bounds. If we again consider the dependence structure between cells, we can construct a global DR measure. Theorem 1 shows that the distribution of the solutions (which are multivariate with respect to the cells) is uniform. If we incorporate this into the reciprocal of the conditional entropy,[3] we

obtain a global measure of DR:

$$\text{DR}_{\text{global}} = 1/\log_2(m_s), \tag{15}$$

where m_s is the number of solutions. The global measure can be used to assess overall risk, whereas the cell-level measures can give upper bounds on the risk associated with a particular cell. We compare our COM-Poisson approach with an interval-bounds-based approach in the following section.

4. Examples

In this section, we try to reconstruct the count tables for three real data sets with different sizes of count tables. Our objective is to illustrate the extent of the risk posed by disclosing the sufficient statistics of the model. In all cases, we use only the sufficient statistics n, s_1, and e^{s_2} for the reconstruction and show the resulting number of solutions, and where possible, the solutions themselves. When the solutions are not available, we give interval bounds for the cells.

4.1. Example 1: 2004 Federal Disaster Declarations

We start with a small frequency table that describes the frequency of federal disaster declarations in 2004 in the 50 U.S. states + the District of Columbia. The data are taken from the official website of the Federal Emergency Management Agency, which belongs to the U.S. Department of Homeland Security and is displayed in Table 3. Fitting a COM-Poisson, we get the estimates $\hat{\lambda} = 1.15$, $\hat{\nu} = 0.88$, which means that the data are over-dispersed compared to an ordinary Poisson distribution. Now, if instead of the complete count table we had only the summary statistics $n = 51$, $s_1 = 61$, and $e^{s_2} = 2^{20} \cdot 3^7$, how many count tables are consistent with these sufficient statistics? In this case, we can bound J by 5, because 5 is the smallest prime number that does not occur in the factorization. We also have $n(2) = 20$ and $n(3) = 7$. To find the actual counts, we compute $a(k, p)$ for $p = 2, 3$ and $k = 2, 3, 4$. From (9) or Table 2, we get $a(2, 2) = a(3, 3) = 1$, $a(4, 2) = 2$, and all other as are zero. This is independent of the data. We now plug these values into (12) and obtain the equations

$$20 = f_2 + f_3 + 3f_4, \tag{16}$$

$$7 = f_3 + f_4. \tag{17}$$

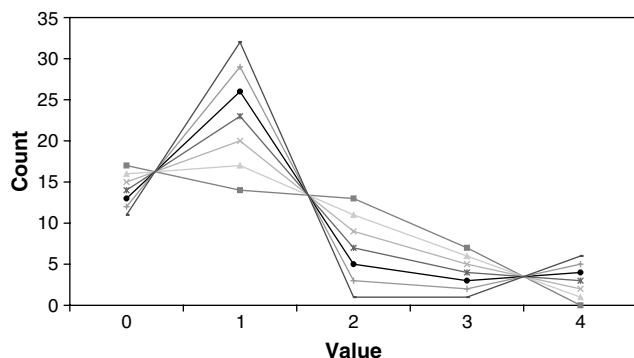We write f_2 and f_3 as a function of f_4, and add the two constraints from 13:

$$51 = f_0 + f_1 + f_2 + f_3 + f_4, \tag{18}$$

$$61 = f_1 + 2f_2 + 3f_3 + 4f_4. \tag{19}$$

Table 3   Number of Federal Disaster Declarations in 2004, in the 50 U.S. States and D.C. ($J = 4$)

No. of disasters	0	1	2	3	4
Frequency	15	20	9	5	2

---

[3] Using the notation from §3.2, this is given by $-1/\sum_{\tilde{f}} P(f \mid S) \cdot \log_2 P(f \mid S) = -1/\sum_{\tilde{f}} (1/m_s) \log_2(1/m_s) = 1/\log_2(m_s)$.

Figure 1 Seven Solutions for Number of Disasters per State

Note. The seven integer solutions of the four equations linking the known sufficient statistics and the unknown counts.

The space of solutions includes seven solutions, each with probability 1/7. These solutions can be easily obtained using software (e.g., Excel) or by hand, and are presented in Figure 1. The largest uncertainty is for counts of one, which range between 14 and 32. Here and in general, if $J = 3$ is known, then the single solution would be $\{f_0 = 17, f_1 = 14, f_2 = 13, f_3 = 7, f_4 = 0\}$. In conclusion, here we disclose more than the information that is required for estimating the parameter in an ordinary Poisson fit, namely, $n(2)$ and $n(3)$. Even with this information at hand, there is only a chance of 1/7 to reconstruct the exact count table, but the seven solutions are easily found. The seven solutions lead to count tables that diverge especially for counts of one. In addition, there are three "shapes" of solutions: one where the counts are decreasing from 0–4, one with a peak at one and then decreasing, and another with a peak at two followed by a decrease and a slight increase at the count of four. Using the interval bounds, we obtain an overall disclosure risk of $DR = 1/\log_2(7) = 0.356$. Cell-level disclosure risks are given in Table 4.

4.2. Example 2: Word Lengths in a Slovak Poem

The count data in Table 5 describe the word lengths in a Slovak poem by M. Rufus, and were used by Wimmer et al. (1994) to describe the distribution of word lengths in texts where is there no restrictions on word length. The authors show that the data follow a COM-Poisson distribution.

Suppose that the complete table was not disclosed, but only the sufficient statistics $n = 117$, $s_1 = 338$, and

Table 4 Interval Bounds for Cells and Their Associated Disclosure-Risk Measure

Number of disasters	0	1	2	3	4
Lower bound	11	14	1	1	0
Upper bound	17	32	13	7	6
DR	0.39	0.24	0.28	0.39	0.39

Note. The global measure here is $DR_{global} = 1/\log_2(7) = 0.356$.

Table 5 Word Lengths (Number of Syllables) of 117 Words in a Slovak Poem ($J = 5$)

Word length	0	1	2	3	4	5
Frequency	0	7	33	49	22	6

$e^{s_2} = 2^{166} \cdot 3^{77} \cdot 5^6$. We can bound J by $J < 7$ because 7 is the smallest prime number that does not appear in the factorization. We therefore have $n(2) = 166$, $n(3) = 77$, and $n(5) = 6$. The five equations are

$$166 = f_2 + f_3 + 3f_4 + 3f_5 + 4f_6, \tag{20}$$

$$77 = f_3 + f_4 + f_5 + 2f_6, \tag{21}$$

$$6 = f_5 + f_6, \tag{22}$$

$$117 = f_0 + f_1 + f_2 + f_3 + f_4 + f_5 + f_6, \tag{23}$$

$$338 = f_1 + 2f_2 + 3f_3 + 4f_4 + 5f_5 + 6f_6. \tag{24}$$

In this case, there are 14 solutions, displayed in Figure 2. It can be seen that the main uncertainty is for counts of two (ranging 33–45). Also, there are two possible "shapes" of counts: one with decreasing counts as the words get longer, and the other with an increase at words with six syllables. If we consider the practical meaning of a word length, thereby eliminating solutions with nonzero counts of 0, the number of valid solutions is seven. In this "small" example $J = 5$ (there are only word lengths of 0–5 words), yet there are 14 possible solutions that have different "shapes." Because the number of solutions is small, it is feasible to compute them. For sake of comparison, we give the interval bounds and their associated DR measures in Table 6.

4.3. Example 3: Injuries in Accidents

The next example considers $J = 11$ and shows how the rate of complexity quickly increases. Table 7 describes the number of injuries from automobile accidents in a sample of 10,000 accidents in the United States in 2001 (from the Department of Transportation's website at

Figure 2 14 Solutions for Word Lengths in a Slovak Poem
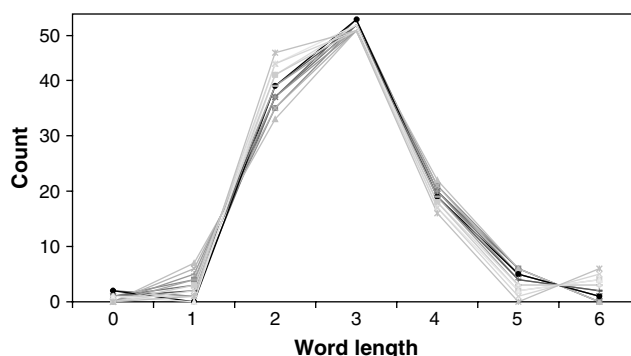
Note. The 14 integer solutions of the four equations linking the known sufficient statistics and the unknown counts.

Table 6 Interval Bounds for Cells and Their Associated Disclosure-Risk Measure

Word length	0	1	2	3	4	5	6
Lower bound	0	0	33	49	16	0	0
Upper bound	2	7	45	51	22	6	6
DR	1	0.36	0.28	1	0.39	0.39	0.39

Note. The global measure here is $DR_{global} = 1/\log_2(14) = 0.263$.

http://www.transtats.bts.gov). The user is given the sufficient statistics: $n = 10{,}000$, $s_1 = 7{,}073$, and $e^{s_2} = 2^{2{,}000} \cdot 3^{585} \cdot 5^{87} \cdot 7^{20} \cdot 11^1$. In this case, the smallest prime number that does not appear in the factorization is $p' = 13$. We therefore have the following seven equations with 13 unknown frequencies:

$$2{,}000 = f_2 + f_3 + 3f_4 + 3f_5 + 4f_6 + 4f_7$$
$$+ 7f_8 + 7f_9 + 8f_{10} + 8f_{11} + 10f_{12}, \qquad (25)$$
$$585 = f_3 + f_4 + f_5 + 2f_6 + 2f_7 + 2f_8 + 4f_9$$
$$+ 4f_{10} + 4f_{11} + 5f_{12},$$
$$87 = f_5 + f_6 + f_7 + f_8 + f_9 + 2f_{10} + 2f_{11} + 2f_{12},$$
$$20 = f_7 + f_8 + f_9 + f_{10} + f_{11} + f_{12},$$
$$1 = f_{11} + f_{12},$$
$$10{,}000 = f_0 + f_1 + f_2 + f_3 + f_4 + f_5$$
$$+ f_6 + f_7 + f_8 + f_9 + f_{10} + f_{11} + f_{12},$$
$$7{,}073 = f_1 + 2f_2 + 3f_3 + 4f_4 + 5f_5 + 6f_6 + 7f_7$$
$$+ 8f_8 + 9f_9 + 10f_{10} + 11f_{11} + 12f_{12}.$$

Using integer programming software (Latte; see http://www.math.ucdavis.edu/~latte/), we find that the number of integer solutions is 82,938,779. Clearly, it is not feasible to find all solutions and therefore it is even hard to compare the different solutions. We therefore resort to computing interval bounds for the cells, with their associated DR measures, in Table 8.

5. Managerial Implications and Future Research

The three examples illustrate the low risk associated with disclosing the COM-Poisson sufficient statistics. Yet, in all three examples, the COM-Poisson well approximates the data distribution, and therefore the

Table 7 Number of Injuries in 10,000 Car Accidents in 2001

Number of injuries	0	1	2	3	4	5	6	7	8	9	10	11
Frequency	5,363	3,091	1,008	348	105	46	19	9	7	2	1	1

sufficient statistics can be used for statistical estimation and decision making. The disclosure risk is low with even small count tables where the solutions can be computed. It becomes extremely low in larger count tables (e.g., with data values that range from 0–11), where it is practically infeasible to compute the solutions. This policy masks not only the actual cell counts but also the size of the table. At the same time as noted, data utility is not sacrificed, thus meeting the needs of legitimate users.

The proposed policy described in this paper offers data owners the ability to control information disclosure taking into account the competing needs of legitimate data users and malicious data snoopers. Sufficient statistics meet the needs of legitimate data users. However, the format in which the sufficient statistic is disclosed (for example, disclosing S_2 in exponentiated form versus in a decimal approximation) has a greater impact on the malicious snooper than it does on the legitimate data user.

Unlike many disclosure limitation techniques that are aimed at categorical or continuous data, our method is suitable for discrete count data. In particular, we focus on one-way tables of discrete count data. An important next step is to develop ways for dealing with higher-dimensional tables, which are prevalent in federal applications. One first needs to define the nature of the multiway data structure. A popular structure is a multiway table where one dimension corresponds to a discrete variable (such as the distribution of the number of abortions per woman, as reported by a survey), and the other dimensions are categorical variables (such as ethnicity, marital status, religion, and state). Consider the schematic example in Table 9. As in the one-way case, the goal is to disclose only enough information to allow for inference about abortions in different states, but the actual cell counts should remain unrevealed. One approach for using the COM-Poisson distribution for disclosure limitation is to treat each category (e.g., state) separately, and report a set of sufficient statistics for each category. If the distribution of abortions in each state

Table 8 Interval Bounds for Cells and Their Associated Disclosure-Risk Measure

Number of injuries	0	1	2	3	4	5	6	7	8	9	10	11	12
Lower bound	4,994	2,686	230	0	0	0	0	0	0	0	0	0	0
Upper bound	5,510	4,213	1,241	477	477	66	66	19	19	19	19	1	1
DR	0.11	0.09	0.10	0.11	0.11	0.17	0.17	0.24	0.24	0.24	0.24	—	—

Note. The global measure here is $DR_{global} = 1/\log_2(82938779) = 0.04$.

Table 9 **Distribution of the Number of Abortions per Woman by State (Schematic Two-Way Table)**

	No. of abortions per woman				
	0	1	2	3	...
Alabama					
Alaska					
Arizona					
⋮					

is independent of the distribution in the other states, then there is no loss of information by treating each state separately. Even if there is a dependence, but the user is only interested in category-level inference, then this policy is sufficient.

If there is a dependence between the discrete variable and the categorical variable(s), and the user is interested in estimating this relationship, then two possible approaches are (1) to supplement the category-level information with some measure of dependence, or (2) to develop a COM-Poisson regression model (perhaps one that generalizes a Poisson regression model) where the predictors are dummy variables that correspond to the categorical information. In the last setting, it is possible to integrate not only categorical dimensions but also continuous ones. These are two avenues that require further research. Finally, a third direction would be to extend the COM-Poisson distribution to a multivariate distribution, which would be useful for multiway tables where there are multiple dimensions that are based on discrete count variables.

Acknowledgments

References

Boyens, C., R. Krishnan, R. Padman. 2004. On privacy-preserving access to distributed heterogeneous healthcare information. *Proc. 37th Annual Hawaii Internat. Conf. on System Sciences* (*HICSS'04*). IEEE Computer Society, Los Alamitos, CA.

Burridge, J. 2003. Information preserving statistical obfuscation. *Statist. Comput.* **13** 321–327.

Chowdhury, S. D., G. T. Duncan, R. Krishnan, S. F. Roehrig, S. Mukherjee. 1999. Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. *Management Sci.* **45**(12) 1710–1723.

Dobra, A., S. E. Fienberg, M. Trottini. 2003. Assessing the risk of disclosure of confidential categorical data. J. Bernardo, ed. *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics.* Oxford University Press, Oxford, UK.

Domingo-Ferrer, J., V. Torra. 2005. Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. *Data Mining Knowledge Discovery.* **22**(11) 195–212.

Domingo-Ferrer, J., A. Oganian, V. Torra. 2002. Information-theoretic disclosure risk measures in statistical disclosure control of tabular data. *Proc. 14th Internat. Conf. on Scientific and Statistical Database Management (SSDBM '02).* IEEE Computer Society, Los Alamitos, CA.

Duncan, G. T., S. E. Fienberg, R. Krishnan, R. Padman, S. F. Roehrig. 2001. Disclosure limitation methods and information loss for tabular data. Pat Doyle, Julia I. Lane, J. M. Theeuwes, and Laura V. Zayatz, eds. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* North-Holland, Amsterdam, The Netherlands, 135–166.

Federal Emergency Management Agency. Retrieved July 15, 2005, http://www.fema.gov.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. John Wiley and Sons, New York.

Fienberg, S. E., U. E. Makov, R. J. Steele. 1998. Disclosure limitation using perturbation and related methods for categorical data. *J. Official Statist.* **14**(4) 485–502.

Garfinkel, R., R. Gopal, P. Goes. 2002. Privacy protection of binary confidential data against deterministic, stochastic, and insider threat. *Management Sci.* **48**(6) 749–764.

Gonzalez, J. F., Jr., L. H. Cox. 2005. Software for tabular data protection. *Statist. in Medicine* **24**(4) 659–669.

Kadane, J. B., G. Shmueli, T. P. Minka, S. Borle, P. Boatwright. 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. *Bayesian Anal.* **1**(2) 363–374.

Li, Y., L. Wang, L. Jajodia. 2002. Preventing interval-based inference by random data perturbation. *Lecture Notes in Computer Science*, Vol. 2482. Springer-Verlag, Berlin, Germany, 160–170.

Sarathy, R., K. Muralidhar. 2002. The security of confidential numerical data in databases. *Inform. Systems Res.* **13**(4) 389–403.

Shmueli, G., T. P. Minka, J. B. Kadane, S. Borle, P. Boatwright. 2005. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *J. Royal Statist. Soc.* **54**(1) 127–142.

Sweeney, L. 2002a. *K*-anonymity: A model for protecting privacy. *Internat. J. Uncertainty, Fuzziness Knowledge-Based Systems* **10**(5) 557–570.

Sweeney, L. 2002b. Achieving *K*-anonymity privacy protection using generalization and suppression. *Internat. J. Uncertainty, Fuzziness Knowledge-Based Systems* **10**(5) 571–588.

Wimmer, G., R. Kohler, R. Grotjahn, G. Altmann. 1994. Toward a theory of word length distributions. *J. Quant. Linguistics* **1** 98–106.