1352.0.55.083

Australian Bureau of Statistics

Research Paper

Measuring and Correcting for Information Loss in Confidentialised Census Counts

www.abs.gov.au

New Issue

Research Paper

Measuring and Correcting for Information Loss in Confidentialised Census Counts

Janice Wooton

Statistical Services Branch

Methodology Advisory Committee

17 November 2006, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 21 JUNE 2007

ABS Catalogue no. 1352.0.55.083
ISBN 978 0 64248 339 3

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Janice Wooton, Statistical Services Branch on Canberra (02) 6252 5764 or email <methodology@abs.gov.au>.

CONTENTS

MEASURING AND CORRECTING FOR INFORMATION LOSS IN CONFIDENTIALISED CENSUS COUNTS

Janice Wooton
Statistical Services

ABSTRACT

The Australian Bureau of Statistics (ABS) has developed a new confidentiality protection method for census tables, to be applied for 2006 census data. The method differs from more traditional disclosure control methods in that there are a number of parameters that can be set to fine-tune the methodology. In order to determine the best settings for these parameters, the ABS has investigated and attempted to balance the benefit (level of protection or reduction of risk of identification) with the cost (damage done to the integrity of the table, or the information loss). This paper discusses a number of ways to measure information loss in tables . In particular, a detailed examination of the χ^2 test of association in a three dimensional table is undertaken. We show that the confidentiality procedure produces a positive bias on this statistic and on certain partitions of it. As a result of this work, we are able to quantify the impact on the χ^2 test of association due to the confidentiality protection, and provide advice for users on how to compensate for this effect.

Keywords: Cell Perturbation, Information Loss Measures, Frequency Tables, Chi-Squared Test of Association, Confidentiality, Statistical Disclosure Control.

1. INTRODUCTION

Tables of counts from the *Census of Population and Housing* are one of the ABS's most widely used products and the ABS is under a legislative obligation to maintain the confidentiality of the people who provide these data. For the 2006 census tabular output there will be improvements in the way users can access census data and finer level geographical building blocks called mesh blocks will be made available. Mesh blocks and improved access modes give users more flexibility and more scope to build and define their own tables. The method of confidentialising census tables used prior to 2006 is no longer adequate in this new environment because the disclosure risk, through table differencing, is too high (for further details see Wooton and Fraser, 2005). The old method adjusts small cells only and tables that differ slightly in their definition could be differenced from one another to obtain detailed unconfidentialised and potentially identifying small subpopulation data.

To solve this problem the ABS has recently developed a new cell perturbation confidentiality methodology to be applied to all census tables of counts before release. This new method protects against disclosures occurring through table differencing because small noise terms are added to all cells and not just to the smaller cells. Therefore under the new method, if a user differences two tables and obtains small cell counts they cannot be certain of the exact original cell values of these small differences thereby protecting small cell counts from being revealed with certainty.

There are various parameters associated with this new cell perturbation method. The parameter values ultimately control both the amount of information loss and the identification risk in a perturbed table. Parameters need to be chosen to give a good compromise between minimising these two conflicting attributes. Before making a decision about the parameter values it will be necessary to somehow measure both information loss and identification risk in the perturbed tables.

The main focus of this paper will be on how to measure the information loss in perturbed census tables of counts. In Section 2 we briefly describe the new cell perturbation methodology and the parameters that need to be chosen. A Monte Carlo study is then undertaken in Sections 3 to 6 which examines: (i) the perturbation distributions and the distortion to the original cell counts; (ii) the variance and covariance structure of the perturbations within tables; and (iii) the impact of the perturbations on contingency table analyses and tests. From this empirical investigation we obtain many useful insights into both information loss and identification risk. We then discuss information loss measures in more detail in Section 7 and how to adjust analyses to correct for the effects of perturbation. These results will also help determine a good choice of perturbation parameter values.

2. BRIEF DESCRIPTION OF THE CELL PERTURBATION METHODOLOGY

Perturbations are added to all cells in a two stage process, described in detail below. For the i-th cell in a table we have (future references drop the i subscript):

$$\begin{aligned} A_i &= U_i + (P_i - U_i) + (A_i - P_i) \\ &= U_i + e_{p(i)} + e_{a(i)} \\ &= U_i + e_{t(i)}, \end{aligned} \tag{1}$$

where A_i is the additively perturbed cell count that the ABS will be publishing, U_i is the original (unconfidentialised) cell count and P_i is the consistently perturbed cell count. There are two random noise terms $e_{p(i)}$ and $e_{a(i)}$ which are, respectively, the discrete stage 1 and stage 2 perturbations. We will now describe how these are generated and why a two stage process is needed.

Before any tables are perturbed or even produced we first assign to each unit on the microdata file a permanent independent discrete random uniform number on the interval $[0, N-1]$, where N is a large positive integer. These are called Rkeys and are used to generate consistent values of e_p. That is, whenever the same group of units are in a cell, the same value of e_p is always generated. This is achieved through a function that maps the contributing Rkeys to a new integer value in the interval $[0, N-1]$. This new value is referred to as the Ckey of the cell. Addition modulo N is one example of a suitable function. Each Ckey and U value are then both mapped to a probability distribution guaranteeing the same random perturbation e_p is always applied whenever the same set of contributors are in a cell.

The distribution for e_p is chosen to balance measures of both information loss and identification risk. Identification risk is related to uncertainty. The more uncertain we are about an outcome, the smaller the identification risk. Information entropy is a measure of the uncertainty of an outcome (see chapter 11 of Jaynes (2003) for further details). Conditional on U, an appropriate distribution of e_p can be obtained by maximising the entropy (uncertainty) subject to some information loss constraints. That is, for each U we maximise the function

$$-\sum_k P(e_p = k \mid U) \log P(e_p = k \mid U)$$

subject to the following constraints:

1. $\sum_k P(e_p = k \mid U) = 1$ and $P(e_p = k \mid U) \geq 0$ for all k.
2. $E(e_p \mid U) = 0$ and $Var(e_p \mid U) = c_U$, where c_U is a non-negative constant that needs to be set.
3. $U + e_p \geq 0$.

4. Given U, $e_p \in [-d_U, -d_U+1, \ldots, -1, 0, 1, \ldots, d_U-1, d_U]$ excluding values in this integer range where 3. above is not satisfied. d_U is a non-negative integer constant that needs to be set.

The parameters values c_U and d_U need to be chosen before any of the e_p are generated. By adjusting these we have some control over both information loss and identification risk.

There is no closed form solution to this problem, but it can be written in terms of Lagrangian multipliers and then solved numerically. The solution is,

$$P(e_p = k \mid U) = e^{\lambda_U - 1 + \alpha_U k + \beta_U k^2},$$

where λ_U, α_U and β_U are chosen to satisfy

$$\sum_k P(e_p = k \mid U) = 1$$
$$\sum_k P(e_p = k \mid U) k = 0$$
$$\sum_k P(e_p = k \mid U) k^2 = c_U$$

An e_p value is generated independently in every cell of the table including marginal and grand total cells. The table defined by the set of $U + e_p$ values is not additive in general (by additive we mean additive relationships such as row totals adding to the grand total). To restore additivity to the table, we add in the second stage perturbation e_a to each cell. Ideally we would like the set of e_a terms for a table to all be as close to 0 as possible to ensure some consistency for the same cells in different tables. To generate the set of e_a for a given table, we use an iterative fitting algorithm developed by the ABS. This algorithm attempts to balance and minimise squared distances to the set of $P = U + e_p$ values for a given table subject to all the additive relationships being maintained between cells. $e_a = 0$ is always guaranteed in grand total cells to ensure consistency of grand totals across tables.

As we have seen, the perturbation process adds random noise values $e_t = e_p + e_a$ to all cells in a table. This means that original cell values and original cell proportions will get distorted under perturbation leading to information loss. In order to determine the amount of distortion it will be necessary to examine the distributional properties of the e_t terms. This examination can only be done empirically via simulation because it is not immediately clear what the exact distributions of the e_t terms will be. In addition it is expected that these distributions will depend on factors such as the number of additivity constraints in a table, the dimension of the table, the number of categories within a dimension, sparsity and the total sample size in a table. A Monte Carlo study is therefore undertaken and the results are discussed in the next section.

3. EXAMINATION OF THE DISTRIBUTIONAL PROPERTIES OF e_t

Information loss is something that needs to be examined from many perspectives. There is no one measure which can sufficiently cover everything we want. However we can assert that when information loss is small, then $Var(e_t)$ should be small, $E(e_t) \approx 0$ (to avoid bias) and $P(|e_t| > d_U)$ should be very small. The shape of the distribution will also play a role and needs to be considered. It is therefore useful to look at these attributes.

Before we simulate any perturbed tables we need to choose the set of perturbation parameters $\{c_U, d_U\}$. For the simulation study these will be $c_0=0$, $d_0=0$; $c_1=2.5$, $d_1=3$; $c_2=4$, $d_2=3$; and for $U \geq 3$, $c_U=4$, $d_U=5$. These were thought to be reasonable initial choices and hopefully would give a good compromise between information loss and disclosure risk.

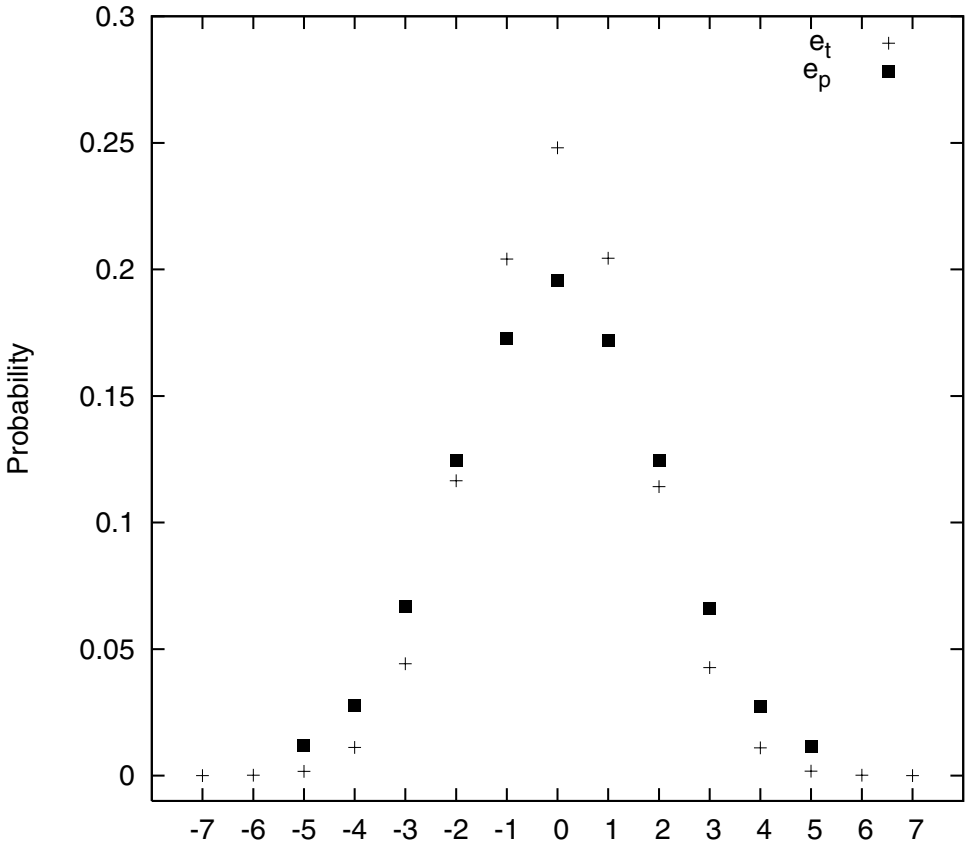
The Monte Carlo study is restricted to the examination of three dimensional weekly individual income by age group by sex tables from the 2001 census. These are the largest tables in terms of the number of cells in the 2001 *Basic Community Profile* series published by the ABS. Two different 2001 SLA (Statistical Local Area) geographies are chosen for the analysis, call these SLA 1 and SLA 2. The SLA 1 table has an average original cell count of 4.7 (it contains about 1500 people) and SLA 2 has an average original cell count of 0.4 (it contains about 100 people). SLA 2 is very sparse and all but one of its interior cells has an original count of 0, 1 or 2.

For each of the two SLAs we simulate 10,000 independent sets of additively perturbed tables. This is done by simulating 10,000 sets of Rkeys for each SLA and then for each set of Rkeys calculating the additively perturbed table using the steps outlined in Section 2 to obtain 10,000 sets of $U + e_p + e_a$ values for each SLA.

For the analysis we group various cells together. The distribution of e_p given U are the same for $U \geq 8$ and is symmetric and bell shaped. Therefore as U gets sufficiently far from 0 we expect e_a and hence e_t to behave the same as well. We group cells with $U \geq 11$ together and the rest we examine individually. There are also various different cell types in a table which are defined by whether the cell is a particular marginal total, subtotal, grand total or interior cell. For our three dimensional cross classified table there are eight different cell types. These cell types can be obtained by summing over particular combinations of dimensions in the table and cells will be grouped according to the cell type.

Figure 3.1 compares the e_p and e_t distributions for interior cells with $U \geq 11$ for SLA 1. The $P(|e_t| > 5)$ is very small and the two distributions are similar in shape with e_t being the more peaked of the two. In general when examining figure 3.1 and other graphs (not given here) for the different cell types and U values, the e_t distributions given U appear to compromise well between information loss and identification risk.

3.1 Comparison of e_p and e_t distributions for the interior cells of SLA 1 (for U $\geq$ 11)

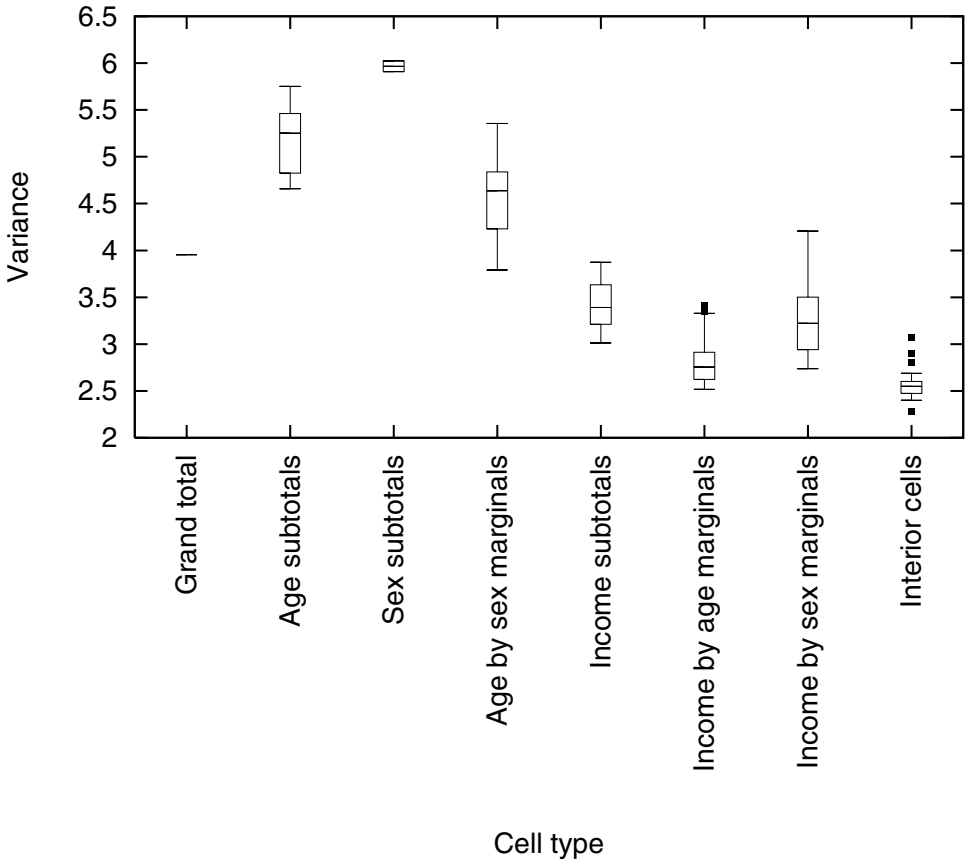


There is always sufficient uncertainty in outcomes and mostly only small noise terms are being added with high probability. Therefore on average the cell count distortions will not be too large.

An estimate of $E(e_t)$ for each cell was calculated based on the 10,000 sets of e_t values. Expectations were all found to be less than 0.5 in magnitude and mostly very close to 0. Therefore $E(e_t) \approx 0$ is a fair conclusion and in general the perturbation process is approximately unbiased. An examination of the estimated cell variances was also undertaken. These were in general found not to be too large.

Figure 3.2 contains boxplots of the distributions of estimated variances of e_t within each cell type for $U \geq 11$ for SLA 1. Notice that the variances of e_t are in general smallest for the interior cells of the table and were largest for certain marginal subtotals. More noise is therefore being introduced to the marginals, although in general the noise is smaller as a proportion of the cell frequency U than is the case for interior cells. The noise in the marginals is also small relative to the noise introduced by some alternative methods. Suppose we had applied uncontrolled random rounding to base 3 to cell counts less than 3 instead. Then the variances in the sex subtotal cells would have been about 55.

3.2 Distribution of e_t cell variances within each cell type



We do have to be careful that the variances are not too small because then we may have too large an identification risk. We also examined correlations between e_p and e_a within each cell and found that for $U \geq 11$ for SLA 1, the interior cells had the largest negative correlations (approximately –0.65). This may not be ideal from an identification risk perspective. This is because we know the value of $U + e_p + e_a$, since this is the published interior cell count say. It is relatively easy to derive $U + e_p$ because we could just request another table with the interior cell count as the grand total. $e_a = 0$ is always guaranteed for grand totals and therefore we can derive e_a for the interior cell via differencing. So given e_a, can we predict e_p with good precision and hence U? If the answer is yes we have a high identification risk. We examined plots of e_p versus e_a and determined that e_p cannot be predicted with good precision for any value. The effect is that in general given e_a, we are roughly halving the variance of e_p.

4. ANALYSIS OF CONTINGENCY TABLES

So far we have only looked at the distributional properties of the perturbations. Now we will determine how perturbations affect contingency table analyses and tests. When analysing contingency tables, we are often interested in answering such questions as:

1. Is there an association between certain categorical variables?
2. What is the nature (or direction) of the association (assuming 1. holds)? For example, can we conclude that income increases with age?

Pearson's χ^2 test and log-linear model analyses can be used to address the above questions. As Beh and Davy (2004) suggest, from the analysis of data using log-linear models, the researcher can determine important associations that exist in the data. However as Beh and Davy (2004) go on to state, there are some problems with this method. One is that the selection of an optimal log-linear model requires a trial and error approach of fitting and refitting which could lead to computing a large number of models. This is not ideal for our situation where we would like to analyse thousands of simulated tables. Some other issues are that the conventional method of estimating parameters is to use an iterative maximum likelihood technique such as Newton–Raphson. To apply this to thousands of tables will be computationally intensive. Also, sometimes the Newton–Raphson procedure may not converge to a solution.

Applying Pearson's χ^2 test has a distinct advantage over the use of log-linear models for addressing question 1. It is computationally easy to calculate, does not involve using iterative methods, can easily be applied to thousands of tables and only needs to be applied once irrespective of the relationship. However, Pearson's χ^2 statistic cannot be used on its own to address question 2. It does not give us any information about the nature of the relationship between variables if there is one. But if at least one of the categorical variables is ordinal, then we can partition Pearson's χ^2 statistic using orthogonal polynomials as outlined in Rayner and Best (2001), Beh and Davy (1999) and Beh and Davy (1998) and undertake more specific directional testing to address question 2.

Beh and Davy (1999) describe a partition which can be used on doubly ordered three way tables. Using this partition, information about the relationship between the variables can be obtained by identifying important associations in terms of the location (linear), dispersion (quadratic) and higher order components. The directions of the associations can also be determined. We will apply this methodology to our income by age by sex tables and use it to determine how associations change after perturbation is applied. Some advantages of this method are that model selection is not necessary, calculating the partitions does not involve iteration and it can easily be

applied to thousands of simulated tables. Interestingly, components of the partitions can also be used to directly estimate parameters in ordinal log-linear models (without iteration). See Beh and Davy (2004) and Beh and Farver (2006) for more details.

5. BRIEF DESCRIPTION OF THE BEH AND DAVY (1999) PARTITIONS OF PEARSON'S χ^2 STATISTIC METHODOLOGY

We have a three way table with a grand total n. The table has I rows ($i = 1, 2, ..., I$), J columns ($j = 1, 2, ..., J$) and K tubes ($k = 1, 2, ..., K$) and the (i, j, k)th cell relative frequency is $p_{ijk} = \frac{n_{ijk}}{n}$.

It is assumed that the rows and columns are ordered and the tubes are not. For our table we take income categories as the rows, age categories as the columns and sex as the tubes. A dot on a subscript indicates summation over that dimension.

Pearson's χ^2 statistic can be partitioned as

$$\begin{aligned}\chi^2 &= \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{uvk}^2 + \sum_{u=1}^{I-1}\sum_{k=1}^{K} Y_{u0k}^2 + \sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{0vk}^2 \\ &= \chi^2_{IJ(K)} + \chi^2_{I(K)} + \chi^2_{J(K)}\end{aligned} \tag{2}$$

where

$$Y_{uvk} = \sqrt{n}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{a_u(i)b_v(j)p_{ijk}}{\sqrt{p_{..k}}}. \tag{3}$$

The set $\{a_u(i)\}$ are orthogonal polynomials on $\{p_{i..}\}$ and the set $\{b_v(j)\}$ are orthogonal polynomials on $\{p_{.j.}\}$. These can be generated using the formulae given on page 70 of Beh and Davy (2004) and we use natural scores. The three χ^2 partitions in (3) each have asymptotic independent χ^2 distributions under the null hypothesis of independence with degrees of freedom $(I-1)(J-1)K$, $(I-1)(K-1)$ and $(J-1)(K-1)$ respectively. The Y_{uvk} values are asymptotically normal with mean 0 under the null hypothesis of independence and can be used to detect any associations on a category level. According to Beh and Davy (1999), Y_{uvk} (for $u > 0$ and $v > 0$) describes the effect the (u,v)th bivariate moment has on the k-th non-ordered tube category, Y_{u0k} describes how the uth univariate moment of the rows affects the k-th non-ordered tube category and Y_{0vk} describes how the vth univariate moment of the columns affects the k-th non-ordered tube category.

The χ^2 partitions in (2) can also be broken down into further partitions as

$$\chi^2_{IJ(K)} = \sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{1vk}^2 + \sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{2vk}^2 + ... + \sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{rvk}^2 + \sum_{u=r+1}^{I-1}\sum_{v=1}^{J-1}\sum_{k=1}^{K} Y_{uvk}^2 \tag{4}$$

$$\chi^2_{IJ(K)} = \sum_{u=1}^{I-1}\sum_{k=1}^{K} Y_{u1k}^2 + \sum_{u=1}^{I-1}\sum_{k=1}^{K} Y_{u2k}^2 + ... + \sum_{u=1}^{I-1}\sum_{k=1}^{K} Y_{urk}^2 + \sum_{u=1}^{I-1}\sum_{v=r+1}^{J-1}\sum_{k=1}^{K} Y_{uvk}^2 \tag{5}$$

$$\chi^2_{I(K)} = \sum_{k=1}^{K} Y_{10k}^2 + \sum_{k=1}^{K} Y_{20k}^2 + ... + \sum_{k=1}^{K} Y_{r0k}^2 + \sum_{u=r+1}^{I-1}\sum_{k=1}^{K} Y_{u0k}^2 \tag{6}$$

$$\chi^2_{J(K)} = \sum_{k=1}^{K} Y^2_{01k} + \sum_{k=1}^{K} Y^2_{02k} + \ldots + \sum_{k=1}^{K} Y^2_{0rk} + \sum_{v=r+1}^{J-1} \sum_{k=1}^{K} Y^2_{0vk} \qquad (7)$$

Each of the $r+1$ partitions on the right hand side of equations (4), (5), (6) and (7) above follow χ^2 distributions asymptotically under the null hypotheses of independence. These can be used for specific directional tests. Common values for choice of r are either 2 or 4. As Rayner and Best (2001) suggest, we are usually only specifically interested in terms relating to the first four moments at most (often two are enough to describe a relationship). A reasonable approach to analysing a contingency table would be to calculate all $r+1$ terms on the right hand side of equations (4) to (7) and do formal tests of significance of these. Once significance is established we could informally look at individual Y_{uvk} values to determine the direction of any associations. This approach is undertaken by both Rayner and Best (2001) and Beh and Davy (1999) and we will do so too.

6. APPLYING PEARSON'S χ^2 PARTITIONS TO THE ORIGINAL AND THE ADDITIVELY PERTURBED SIMULATED TABLES

We have doubly ordered three way tables with $i=1, 2, \ldots, 14$ ordinal income categories (in order of increasing income), $j=1, 2, \ldots, 8$ ordinal age categories (in order of increasing age) and $k=1,2$ non-ordinal sex categories (1=male and 2=female). The analysis is restricted to positive stated income.

SLA 2 is very sparse which means that the asymptotic χ^2 null distribution may not hold. We calculate Monte Carlo p-values for both SLA tables. This is done by conditioning on the income subtotals, sex subtotals and age subtotals as suggested in Mehta and Patel (1997) and then simulating 10,000 tables under the null hypothesis of complete independence using the algorithm in Agresti *et al.* (1979). For each of these 10,000 tables we calculate all the χ^2 partitions whose distributions can then be used to calculate p-values. These distributions will be denoted by 'independence' in future references.

SLA 2 also contains zeros in some of the marginal totals. We add a small value to each cell before calculating any partition. Justification for this can be established through a Bayesian argument (see page 607 of Agresti, 2002). All that is needed is a prior guess of the cell probabilities. We use the Australia level table as our prior guess of the cell probabilities. A constant multiplied by the prior cell probability of the i-th cell is then added to the i-th cell count before the χ^2 statistics are calculated. The constant chosen is 0.1.

We now apply the methodology outlined in Section 5 to both the original SLA tables with $r=4$ (r is defined in Section 5). For SLA 2, Pearson's χ^2 statistic is not significant (p-value=0.5156). For SLA 1, Pearson's χ^2 is highly significant (p-value <0.00001) suggesting that there is an association between income, age and sex. All χ^2 partitions (4) to (7) for SLA 1 are significant overall except for (7). This implies that when income is ignored there is little evidence of an association between age and sex but there is an association when comparing other dimensions. The three largest Y_{uvk} values in terms of their magnitude are $Y_{121} = -6.3$, $Y_{102} = -6.3$ and $Y_{101} = 5.9$. Y_{121} describes the linear by quadratic association between income and age for males and it is negative. Y_{102} and Y_{101} are the two income location terms (ignoring age). The values of these suggest that females tend to have smaller incomes than males. There are other significant terms, but we will not give detailed interpretations here.

In any case, the question we are interested in is to what extent do these conclusions get changed after perturbation? The Y_{uvk} values are important because they determine the magnitude and direction of an association. We calculated all the Y_{uvk} for the simulated additively perturbed tables and a typical distribution is summarised in figure 6.1.

6.1 Distributions of Y_{121} for SLA 1

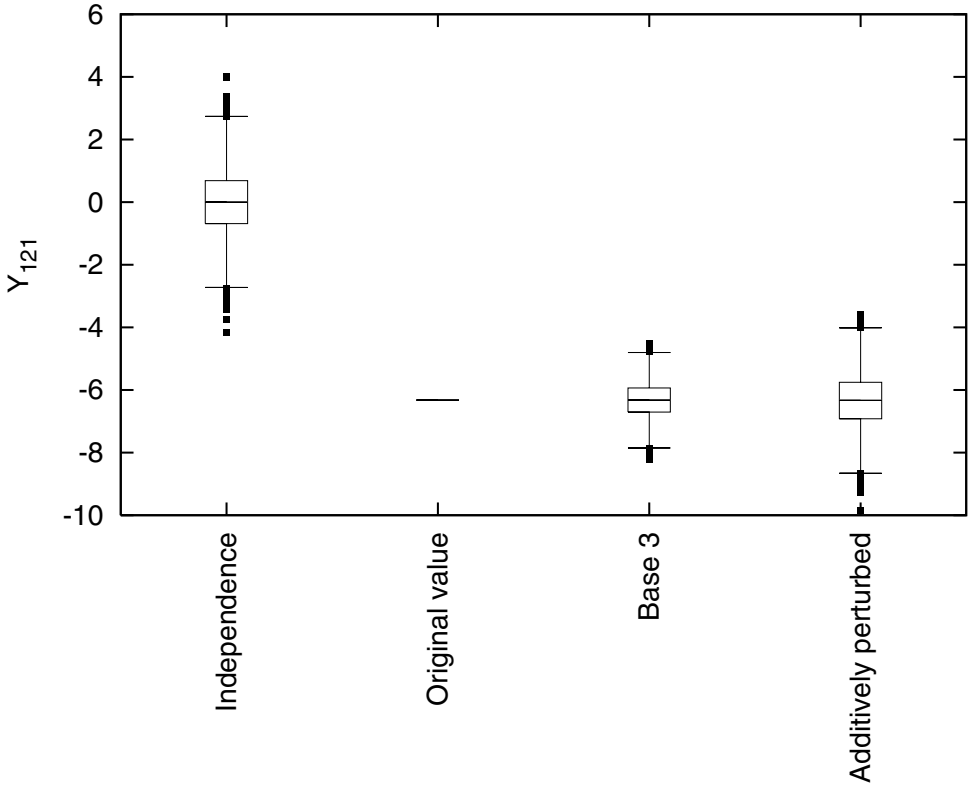
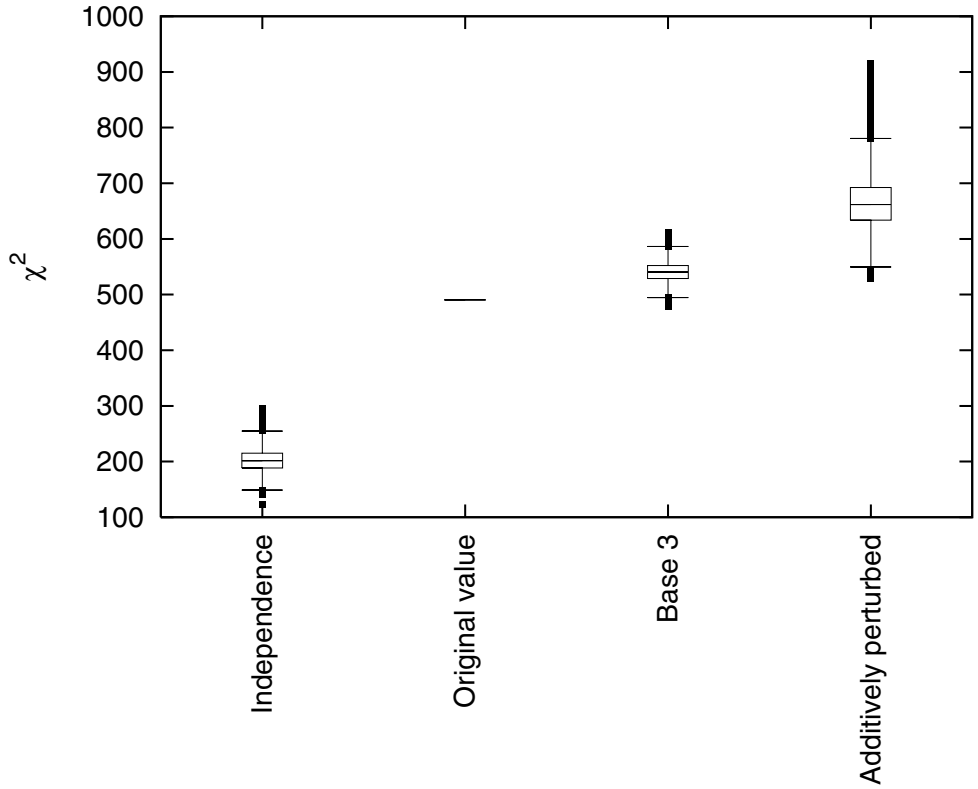


Figure 6.1 shows boxplots of the distributions for component Y_{121} in SLA 1. The 'independence' distribution is the distribution of Y_{121} under the null hypotheses of independence. 'Original value' is the value of Y_{121} for the original unperturbed table. 'Base 3' is the distribution of Y_{121} obtained under an alternative confidentiality method which applies uncontrolled random rounding to base 3 of the 1's and 2's. Base 3 is included for comparative purposes. 'Additively perturbed' is of course the distribution of Y_{121} under perturbation.

We generated many similar plots for the other components as well and a consistent pattern emerged. The distribution of Y_{uvk} under perturbation (and Base 3) is roughly centred around the original Y_{uvk} value. So under perturbation we are adding noise to each component. Denote the noise by e_{uvk} and this noise term follows a roughly symmetric distribution with mean 0. That is, $Y^*_{uvk} = Y_{uvk} + e_{uvk}$, where Y^*_{uvk} is the Y_{uvk} component we obtain under perturbation. To obtain Pearson's χ^2 value and the other partitions described in equations (4) to (7) for the original table, we add appropriate Y_{uvk} sums of squares together and calculate,

$$\text{Partition} = \sum_{m=1}^{M} Y_m^2, \tag{8}$$

6.2 Distributions of Pearson's χ^2 statistic for SLA 1



where m denotes a particular combination of uvk values and M is the total number of squared components we add together.

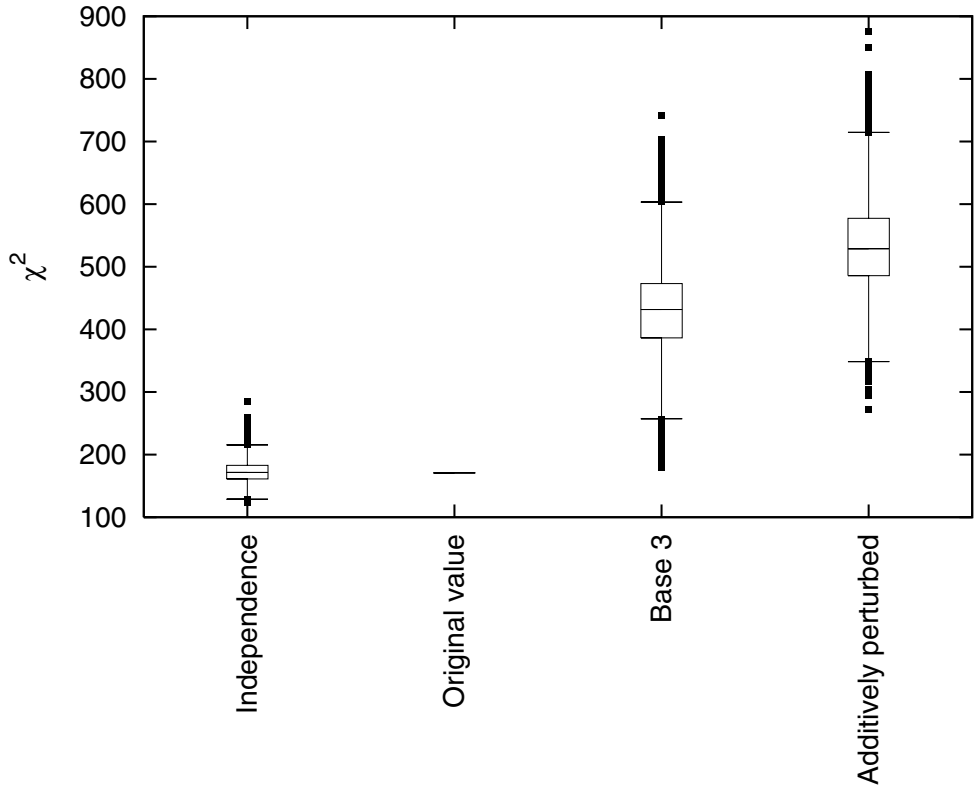
After perturbation is applied to a table we only have the Y^*_m values available and so the partitions are calculated using,

$$\begin{aligned} \text{Partition}^* &= \sum_{m=1}^{M} (Y_m^*)^2 \\ &= \sum_{m=1}^{M} (Y_m^2 + 2Y_m e_m + e_m^2) \end{aligned} \tag{9}$$

If we assume that $E(e_m) = 0$, which is a reasonable assumption as noted above, then the expected partition under perturbation is,

$$\begin{aligned} E(\text{Partition}^*) &= \sum_{m=1}^{M} \left(Y_m^2 + 2Y_m E(e_m) + E(e_m^2) \right) \\ &= \sum_{m=1}^{M} \left(Y_m^2 + Var(e_m) \right) \\ &= \text{Partition} + \sum_{m=1}^{M} Var(e_m) \end{aligned} \tag{10}$$

6.3 Distributions of Pearson's χ^2 statistic for SLA 2



This implies that in general under perturbation, there will be an upward bias on each partition including Pearson's χ^2 statistic. The implication of this is that in general the p-values of the partitions will be smaller than the original table, potentially giving a false impression of significance.

Figures 6.2 and 6.3 contain boxplots of the distributions of Pearson's χ^2 statistics under perturbation for SLA 1 and SLA 2. These graphs clearly show an upward bias in the statistics as expected under both perturbation and random rounding to base 3 of the 1's and 2's. For SLA 2, the bias is much larger and the χ^2 statistics are more variable. For SLA 2, most of the time we would conclude that there is an association when the original value suggested otherwise.

7. INFORMATION LOSS MEASURES

So far we have examined the distributional properties of the cell perturbations and how perturbation affects contingency table analyses. Perturbation leads to some distortion of the original cell counts and introduces an upward bias to Pearson's χ^2 statistic. Ideally users would like to be able to adjust their analyses to account for this damage. It is therefore of interest to develop a set of information loss measures that could potentially be published, informing users of the impact of the confidentiality procedure. One useful reference on this subject is Shlomo and Young (2005). We take a similar approach to these authors and divide the information loss measures according to the statistical aspect to be measured.

We saw in Section 6 that perturbation leads to an upward bias in χ^2 statistics. Although we used a specific χ^2 decomposition in Section 6, relevant for a doubly ordered three way table, we still expect that in general there will be an upward bias in χ^2 statistics in other types of tables as well. One measure of information loss could focus on this bias. That is, calculate

$$E(\chi^{2^*}) - \chi^2 \approx \sum_{m=1}^{K(IJ-1)} Var(e_m), \tag{11}$$

where χ^{2*} is the value of χ^2 under perturbation and e_m is as defined in Section 6. We could then calculate the average percentage increase in χ^2 as,

$$\frac{\sum_{m=1}^{K(IJ-1)} Var(e_m)}{\chi^2} \times 100\%, \tag{12}$$

and (12) can be estimated using

$$\frac{\sum_{m=1}^{K(IJ-1)} Var(e_m)}{\chi^{2^*} - \sum_{m=1}^{K(IJ-1)} Var(e_m)} \times 100\%, \tag{13}$$

assuming $\sum_{m=1}^{K(IJ-1)} Var(e_m)$ is known (or replaced with an estimate) and the denominator is positive.

The measure defined at (13) gives users an idea of the average amount of information loss in percentage terms inherent in the χ^2 test of association for a given perturbed table. The variances of the e_m terms will depend on certain table attributes, such as the number of small cells, the number of categories, dimensions and the number of additivity constraints. It may be possible to build up a model to predict these

variances given certain table properties, instead of relying on simulations which are computationally intensive. This is a topic for further research.

In any case, given an estimate of the e_m variances, the bias in χ^2 statistics can be corrected for. Instead of using χ^{2*}, the user should use

$$\chi^{2*} - \sum_{m=1}^{K(IJ-1)} Var(e_m)$$

instead (assuming this value is positive). There will also be a variance on this difference. If this was known then an approximate conservative confidence interval for Pearson's χ^2statistic under perturbation could be calculated.

As we saw in Section 6, the Y_{uvk} are also important because they determine the magnitude and direction of an association. If users had an idea of the variance of these terms under perturbation (they have already been found to be approximately unbiased), then approximate confidence intervals for these could be calculated as well. Confidence intervals are good information loss measures because the length of these indicates the uncertainty about a parameter or statistic we are introducing due to perturbation.

In contingency table analyses we are treating the census data as though they were a random sample from a superpopulation. We can hypothesise that the data was generated from some parametric model. In Beh and Davy (1999) a model of association is defined for a doubly ordered three way table as,

$$p_{ijk} = p_{i..}p_{.j.}p_{..k}\left[1 + \sum_{u=1}^{I-1}\frac{Y_{u0k}a_u(i)}{\sqrt{np_{..k}}} + \sum_{v=1}^{J-1}\frac{Y_{0vk}b_v(j)}{\sqrt{np_{..k}}} + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\frac{Y_{uvk}a_u(i)b_v(j)}{\sqrt{np_{..k}}}\right], \quad (14)$$

where under the null hypothesis of independence all the Y_{uvk} are 0. When analysing our doubly ordered three way census table we could hypothesise and then assume that the census data were generated from a multinomial model with the above cell probabilities. We are then interested in estimating the set of superpopulation parameters $\{Y_{uvk}\}$ and the sets of nuisance parameters $\{p_{i..}\}$, $\{p_{.j.}\}$ and $\{p_{..k}\}$. The Census data give us information about these parameters. To measure the amount of information about a particular parameter inherent in a sample we could use the expected Fisher information measure. See for example Mathai and Rathie (1975) or Azzalini (1996) for further details. This measure of information is defined as,

$$E\left\{\left(\frac{\partial}{\partial\theta}\log L\right)^2\right\}, \quad (15)$$

where θ is a particular parameter we are interested in estimating and L is the likelihood.

We can use Fisher's expected information to help us derive an information loss measure. Firstly, let's simplify the model defined at (14). Suppose we are interested in estimating the probability of a given population unit being in a particular cell in the contingency table, where the overall sample is fixed to be n. We have a sample size of N_1 in this cell and sample size of $n - N_1$ not in the cell. Denote the probability of being in the cell by p_1. Our aim is to estimate the superpopulation parameter p_1 and for simplicity we will assume here that the cell count N_1 follows a Binomial (n, p_1) distribution. It can be easily proved that the maximum likelihood estimator of p_1 is

$$\hat{p}_1 = \frac{N_1}{n}$$

and the expected Fisher Information is,

$$\frac{1}{Var(\hat{p}_1)}, \tag{16}$$

and (16) can be estimated using

$$\frac{n}{\hat{p}_1(1-\hat{p}_1)}.$$

Under perturbation, we add random noise to each cell. So instead of observing N_1 and n we observe $N_1^* = N_1 + e_{t(1)}$ and $n^* = n + e_{p(s)}$, where $e_{t(1)}$ is the total additive perturbation for the interior cell and $e_{p(s)}$ is the consistent stage 1 perturbation added to the grand total cell ($e_a = 0$ is ensured in grand total cells). See Section 2 for a definition of these perturbations. Because of perturbation we can no longer estimate p_1 with $\hat{p}_1$. Instead we use

$$\hat{p}_1^* = \frac{N_1^*}{n^*}$$

which leads to some information loss and an increase in the variance. To get a measure of the information loss due to perturbation we could calculate the expected Fisher information using the joint probability of N_1^* and n^* and subtract this from (16). That is, we could calculate

$$\frac{1}{Var(\hat{p}_1)} - E\left(\left(\frac{\partial}{\partial p_1}\log P(N_1^*, n^*)\right)^2\right), \tag{17}$$

But (17) is difficult to compute. Instead a rough measure of the amount of information loss due to perturbation can be defined as (in part suggested by the form of (16)),

$$\text{Information loss} = \frac{1}{Var(\hat{p}_1)} - \frac{1}{Var(\hat{p}_1^*)}, \tag{18}$$

with the term $Var(\hat{p}_1^*)$ in (18) approximated using a first-order Taylor series expansion. This approximation is,

$$Var(\hat{p}_1^*) \approx Var(\hat{p}_1) + \frac{Var(e_{t(1)})}{n^2} + \frac{p_1^2 Var(e_{p(s)})}{n^2} - \frac{2p_1 Cov(e_{t(1)}, e_{p(s)})}{n^2}. \tag{19}$$

The $Cov(e_{t(1)}, e_{p(s)})$ term in (19) will be positive in general. An approximate conservative estimator of $Var(\hat{p}_1^*)$ can be found by setting $Cov(e_{t(1)}, e_{p(s)}) = 0$ in (19). Assuming this, we can now calculate an estimate for the information loss using (19) and replacing p_1 with $\hat{p}_1^*$ and n with n^*. This estimate is,

$$\text{Information loss} = \frac{n^*}{\hat{p}_1^*(1-\hat{p}_1^*)} - \frac{n^*}{\hat{p}_1^*(1-\hat{p}_1^*) + \dfrac{\widehat{Var}(e_{t(1)})}{n^*} + \dfrac{\hat{p}_1^{*2}\widehat{Var}(e_{p(s)})}{n^*}}, \tag{20}$$

assuming we have estimates of $Var(e_{t(1)})$ and $Var(e_{p(s)})$ available and $\hat{p}_1^*$ is non-zero (if $\hat{p}_1^* = 0$ we could replace it with a small value instead). We can also estimate the percentage information loss as,

$$\frac{\dfrac{1}{Var(\hat{p}_1)} - \dfrac{1}{Var(\hat{p}_1^*)}}{\dfrac{1}{Var(\hat{p}_1)}} \times 100\% \approx \frac{\dfrac{\widehat{Var}(e_{t(1)})}{n^*} + \dfrac{\hat{p}_1^{*2}\widehat{Var}(e_{p(s)})}{n^*}}{\hat{p}_1^*(1-\hat{p}_1^*) + \dfrac{\widehat{Var}(e_{t(1)})}{n^*} + \dfrac{\hat{p}_1^{*2}\widehat{Var}(e_{p(s)})}{n^*}} \times 100\%. \tag{21}$$

This suggests that information loss with respect to estimating a cell proportion decreases as n increases.

Now suppose that the original cell counts in our table are fixed and not generated from a superpopulation model. Instead of observing the fixed count n_1 in a particular cell we observe $n_1^* = n_1 + e_{t(1)}$ instead. This leads to distortion to the original cell counts. If we had information available about the variance of the e_t terms in each cell, then approximate confidence intervals for the original cell counts could be calculated giving the user an idea of the amount of information loss and uncertainty introduced. This would help users make more informed decisions. A rough measure of the amount of information loss in a particular cell can be approximated using the coefficient of variation (by noting that $E(n_1^*) \approx n_1$),

$$\frac{\sqrt{Var(e_{t(1)})}}{n_1}, \tag{22}$$

or estimated using

$$\frac{\sqrt{\widehat{Var}(e_{t(1)})}}{n_1^*}$$

assuming $n_1^* > 0$ and an estimate of $Var(e_{t(1)})$ is available. If the coefficient of variation is large then the estimate n_1^* of n_1 is considered unreliable and the amount of information loss relative to the cell size is large.

So far we have focused on the distortion to the original cell counts in particular cells. It may also be useful to publish overall measures of cell distortion for a table. Shlomo and Young (2005) describe various distance metrics that can be used for this purpose. For example they apply Hellinger's distance, a relative absolute distance and an average absolute distance metric to measure distances between the original cell frequencies and the perturbed cell frequencies. See page 5 of Shlomo and Young (2005) for further details.

8. CONCLUSION

The ABS' new cell perturbation methodology is designed to minimise information loss in tables subject to certain identification risk constraints. Parameter values associated with the method can be chosen to control to some extent both of these conflicting attributes. In this paper we empirically examined the distributional properties of the cell perturbations for two example tables and gained insights into how cell counts will get distorted under perturbation. We also examined the impact of perturbation on contingency table analyses and tests for these tables.

We demonstrated that perturbation will lead to an upward bias in the Pearson's χ^2 statistic and this result also applies to tables that have been subject to random rounding. This upward bias means that p-values under perturbation will be on average smaller and this may lead to a false impression of significance in χ^2 tests of association. The bias was shown to be the sum of the variances of certain noise terms. So if estimates of these variances were available then users could adjust down the χ^2 statistic accordingly.

As demonstrated by Beh and Davy (1999), components of the partitions of Pearson's χ^2 statistic can be used to get an idea of the magnitude and direction of certain associations in a table on a category level. These associations are approximately unbiased under perturbation and for tables with larger grand totals, the variance of these should be small.

The earlier sections of this paper relied on results from a simulation study of two tables. Future work may involve looking at tables with different numbers of interior cells, grand totals, dimensions and additivity constraints. It may be possible to apply a model which predicts the perturbation variances given certain known table attributes. It would also be useful to look further at the correlation structure of components of the χ^2 statistic under perturbation. This information is needed in order to determine approximate confidence intervals for this statistic and the other smaller partitions.

Finally in this paper we demonstrated that there are a variety of different information loss measures that can be applied to perturbed tables. It was found that in general information loss will decrease as the grand total in a table increases. There is no one ideal measure of information loss and this attribute is often hard to measure. Further work needs to be done in this area to determine the best suite of information loss measures to communicate to users.

REFERENCES

Agresti, Alan; Wackerly, Dennis and Boyett, James M. (1979) "Exact Conditional Tests for Cross-Classifications: Approximation of Attained Significance Levels", *Psychometrika*, 44(1), pp. 75–83.

Agresti, Alan (2002) *Categorical Data Analysis*, Wiley.

Azzalini, Adelchi (1996) *Statistical Inference Based on the Likelihood*, Chapman and Hall, London.

Beh, Eric J. and Davy, Pamela J. (1998) "Partitioning Pearson's Chi-Squared Statistic for a Completely Ordered Three-way Contingency Table", *Australian and New Zealand Journal of Statistics*, 40(4), pp. 465–477.

Beh, Eric J. and Davy, Pamela J. (1999) "Partitioning Pearson's Chi-Squared Statistic for a Partially Ordered Three-way Contingency Table", *Australian and New Zealand Journal of Statistics*, 41(2), pp. 233–246.

Beh, Eric J. and Davy, Pamela J. (2004) "A Non-Iterative Alternative to Ordinal Log-Linear Models", *Journal of Applied Mathematics and Decision Sciences*, 8(2), pp. 67–86.

Beh, Eric J. and Farver, Thomas B. (2006) "An Evaluation of Noniterative Methods for Estimating the Linear-by Linear Parameter of Ordinal Log-Linear Models", currently in review with the *International Statistical Review* journal.

Jaynes, E.T. (2003) *Probability Theory: The Logic of Science*, Cambridge University Press.

Mathai, A.M. and Rathie, P.N. (1975) *Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications*, Wiley, New York.

Mehta, Cyrus R. and Patel, Nitin R. (1997) *Exact Inference for Categorical Data*, unpublished paper, Harvard University and Cytel Software Corporation.

Rayner, J.C.W. and Best, D.J. (2001) *A Contingency Table Approach to Nonparametric Testing*, Chapman and Hall/CRC, Boca Raton.

Shlomo, Natalie and Young, Caroline (2005) *Information Loss Measures for Frequency Tables*, UNECE work session on statistical data confidentiality.

Wooton, Janice and Fraser, Bruce (2005) "A Review of Confidentiality Protections for Statistical Tables", *Methodology Advisory Committee Papers*, cat. no. 1352.0.55.072, Australian Bureau of Statistics, Canberra.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS web site can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

2000001564110
ISBN 9780642483393

RRP $11.00

1352.0.55.083 · RESEARCH PAPER: MEASURING AND CORRECTING FOR INFORMATION LOSS IN CONFIDENTIALISED CENSUS COUNTS (METHODOLOGY ADVISORY COMMITTEE) · November 2006