CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Fair risk-utility comparison of tabular perturbation methods by post-processing to expected frequencies.

Øyvind Langsrud (Statistics Norway)
Oyvind.Langsrud@ssb.no

Abstract

In a Eurostat-granted project within Statistics Norway, the cell key method (CKM) and the small count rounding method (SCR) implemented in the r package SmallCountRounding are to be compared. CKM perturbs all values without additivity constraints. SCR perturbs a number of inner cells so that small frequencies are avoided in the aggregated data to be published. A fair comparison is challenging since the methods are very different.

Inspired by synthetic data methods, one approach is to generate expected inner cell frequencies from the perturbed tables. These expected frequencies can be viewed as the best guess of individual level data, although the frequencies are not whole numbers. Since SCR is additive, expected frequencies can be generated by a variant of iterative proportional fitting (IPF). This estimation coincides with log-linear modelling. This approach is not directly applicable to CKM. The data needs to be additivity restored first and we can perform this by least squares estimation. The results are then maximum likelihood estimates under the assumption of gaussian noise. This is a simplification compared to using the actual discrete noise distribution. A modification is needed to ensure non-negativity. Afterwards, expected inner cell frequencies can be obtained by IPF.

No extra information has been added during this post-processing, but individual disclosive information is easier available. Appropriate measures of risk and utility can be made based on the post-processed data and the original data. We will discuss possible measures that can be used to compare CKM, SCR, as well as other methods. In addition, they can be used as guidance for fine-tuning each method.

Fair risk-utility comparison of tabular perturbation methods by post-processing to expected frequencies

Øyvind Langsrud*, Daniel P. Lupp**

* Statistics Norway, oyvind.langsrud@ssb.no

** Statistics Norway, daniel.lupp@ssb.no

Abstract. In a Eurostat-granted project within Statistics Norway, the cell key method (CKM) and the small count rounding method (SCR) implemented in the R package SmallCountRounding are to be compared. CKM perturbs all values without additivity constraints. SCR perturbs a number of inner cells so that small frequencies are avoided in the aggregated data to be published. A fair comparison is challenging since the methods are very different. Inspired by synthetic data methods, one approach is to generate expected inner cell frequencies from the perturbed tables. These expected frequencies can be viewed as the best guess of individual level data, although the frequencies are not whole numbers. Since SCR is additive, expected frequencies can be generated by a variant of iterative proportional fitting (IPF). This estimation coincides with log-linear modelling. This approach is not directly applicable to CKM. The data needs to have additivity restored first, which we can achieve by least squares estimation. The results are then maximum likelihood estimates under the assumption of Gaussian noise. This is a simplification compared to using the actual discrete noise distribution. A modification is needed to ensure non-negativity. Afterwards, expected inner cell frequencies can be obtained by IPF. No extra information has been added during this post-processing, but individual disclosive information is more easily available. Appropriate measures of risk and utility can be made based on the post-processed data and the original data. In this paper we discuss possible measures that can be used to compare CKM, SCR, as well as other methods. In addition, they can be used as guidance for fine-tuning each method.

1 Introduction

Regardless of whether tabular data or microdata is to be released, statistical disclosure control is about controlling the risk of revealing information about the individual statistical units. In this paper, in order to assess risk we apply a micro data approach to deal with tabular frequency data. More specifically, our approach uses the inner frequency table obtained by crossing all of the main dimensional variables. The microdata underlying the tabular frequency data can always be represented by this inner frequency table.

Table 1: Original inner frequencies

	young		middle		old	
party	male	female	male	female	male	female
A	0	0	8	4	4	1
B	0	1	3	5	1	0
C	2	3	9	6	2	7

Table 2: Original aggregated frequencies

party	young	middle	old	male	female	Total
A	0	12	5	12	5	17
B	1	8	1	4	6	10
C	5	15	9	13	16	29
Total	6	35	15	29	27	56

When the inner frequencies are not directly available, they can be generated by a post processing technique (Langsrud, 2019). Below we compute expected inner frequencies which can be viewed as the best guess of individual level data. The general algorithm involves both iterative proportional fitting (IPF) and least squares estimation. Although the frequencies are not whole numbers, they are well suited as input to general methods for risk and utility calculations.

In the discussions below we consider the three-way $(3 \times 2 \times 2)$ example data given in Table 1. The variables are age, sex, and party affiliation and below we consider the latter variable to be sensitive. The aggregated cells considered to be published are in Table 2. However, a perturbation method is required, and in the next section, the data are perturbed by the cell key method (CKM) based on Thompson et al. (2013) and by the small count rounding method (SCR) described in Langsrud and Heldal (2018) Then, in the following sections, we discuss comparable measures of utility and risk.

2 Perturbation by CKM and SCR

Table 3 contains aggregated frequencies perturbed by CKM. The underlying perturbation table was generated by the R package `ptable` (Enderle, 2021) with parameter setting $D = 5$, $V = 3$ and $js = 2$. The latter parameter specifying that 1's and 2's are avoided. It is easy to see that Table 3 is not additive. The published total for party A is 18, but calculated from the age and sex subtotals it becomes 21 and 16, respectively.

Table 4 contains aggregated frequencies perturbed by SCR using R package `SmallCountRounding` (Langsrud and Heldal, 2021) with 3 as the rounding base. Tables perturbed by SCR are always additive since the method is based on rounding small inner cell frequencies. The rounded inner frequencies underlying Table 4 are given in Table 5. As can be seen, small frequencies (ones and twos)

Table 3: Cell key perturbed frequencies to be published

party	young	middle	old	male	female	Total
A	0	16	5	12	4	18
B	0	10	3	3	4	10
C	5	11	7	10	16	29
Total	5	37	15	31	29	57

Table 4: Small count rounded frequencies to be published

party	young	middle	old	male	female	Total
A	0	12	5	12	5	17
B	3	8	0	3	8	11
C	5	15	9	13	16	29
Total	8	35	14	28	29	57

still occur in Table 5. However, this table is hidden from the user. The aim of the algorithm is to limit rounding to a number of necessary inner frequencies. Since Table 5 is hidden, it cannot be used as a basis for calculating risk and utility.

3 Fixing additivity by least squares

To calculate expected inner frequencies, we will first, when needed, restore additivity. To describe the method, we let the vector $\boldsymbol{y}$ consist of all the elements of the original inner frequency table. Furthermore, we let $\boldsymbol{z}$ be the vector of original aggregated frequencies which can be computed from $\boldsymbol{y}$ via a dummy matrix $\boldsymbol{X}$:

$$\boldsymbol{z} = \boldsymbol{X}^T \boldsymbol{y} \tag{1}$$

When the perturbation method is a noise addition method, we can modify this equation as

$$\boldsymbol{z}_{\text{perturbed}} = \boldsymbol{X}^T \boldsymbol{y} + \text{error} \tag{2}$$

Now we want to find an estimate of $\boldsymbol{z}$ from $\boldsymbol{z}_{\text{perturbed}}$. The specific noise distribution is unknown to the user, but the assumption of Gaussian noise is a reasonable starting point. The maximum likelihood estimates are then obtained by least squares. In this modeling, $\boldsymbol{y}$ consists of the unknown parameters. The equation is over-parameterized so the solution for $\boldsymbol{y}$ is not unique, but a unique solution for $\boldsymbol{z}$ can be found. A problem is that negative elements within the estimated $\boldsymbol{z}$ are possible. Here we use a practical and efficient way to handle this. Any negative estimates are set to zero and the remaining z-values are re-estimated by least squares. The additivity-restored version of Table 3, using this method, is given in Table 6. The values are not whole numbers, but this is not a problem in our approach.

Table 5: Small count rounded inner frequencies

	young		middle		old	
party	male	female	male	female	male	female
A	0	0	8	4	4	1
B	0	3	3	5	0	0
C	2	3	9	6	2	7

Table 6: Additivity-restored cell key perturbed frequencies

party	young	middle	old	male	female	Total
A	0.0000	14.9688	3.9687	13.5937	5.3438	18.9375
B	0.0000	8.8437	1.8438	4.9688	5.7187	10.6875
C	5.8182	12.9119	8.9119	10.9460	16.6960	27.6420
Total	5.8182	36.7244	14.7244	29.5085	27.7585	57.2670

4 Inner cells by iterative proportional fitting

We turn back to equation 1 and consider a situation with $\boldsymbol{z}$ known and with $\boldsymbol{y}$ unknown. To estimate $\boldsymbol{y}$ from $\boldsymbol{z}$ we can make use of log-linear modeling, which is common for count data. The specific model to be used is the log-linear model where $\boldsymbol{z}$ is sufficient, which also means it is a Poisson regression model with $\boldsymbol{X}$ as the matrix of independent variables. Our estimate of $\boldsymbol{y}$ are simply the expected frequencies under this model. This estimate can be found by iterative proportional fitting (IPF), which is a standard approach to fitting log-linear models.

Instead of assuming that $\boldsymbol{z}$ consists of original frequencies, we now consider the additivity-restored (if needed) perturbed data as $\boldsymbol{z}$. Although the starting point for the log-linear modeling is counts, the IPF estimation method does not require that $\boldsymbol{z}$ consist of whole numbers. By using this method, with Table 6 as input $\boldsymbol{z}$, Table 7 shows expected inner cell frequencies based on the cell key perturbed data. Similarly, expected inner cell frequencies based on the small count rounded data are given in Table 8. In this case, Table 4 could be used directly as input $\boldsymbol{z}$, since this table is additive. It is clear that Table 5 (hidden) and Table 8 are different, but both add up to Table 4.

5 Utility

To discuss utility, we focus on a measure based on the Hellinger distance, hereinafter referred to as Hellinger utility (Shlomo et al., 2015), which can be written as

$$\text{utility} = 1 - HD(f, g)/\sqrt{\sum f} \tag{3}$$

where

$$HD(f, g) = \sqrt{\frac{1}{2}\sum\left(\sqrt{f} - \sqrt{g}\right)^2} \tag{4}$$

Table 7: Expected inner cell frequencies from the cell key perturbed table

	young		middle		old	
party	male	female	male	female	male	female
A	0.0000	0.0000	10.7449	4.2239	2.8489	1.1199
B	0.0000	0.0000	4.1116	4.7322	0.8572	0.9866
C	2.3040	3.5142	5.1130	7.7989	3.5291	5.3829

Table 8: Expected inner cell frequencies from the small count rounded table

	young		middle		old	
party	male	female	male	female	male	female
A	0.0000	0.0000	8.4706	3.5294	3.5294	1.4706
B	0.8182	2.1818	2.1818	5.8182	0.0000	0.0000
C	2.2414	2.7586	6.7241	8.2759	4.0345	4.9655

and where f and g are vectors of original and perturbed counts, respectively. This measure is bounded between 0 and 1 and 1 represents maximal utility (same as original data). Hellinger utility is included in the package `SmallCountRounding` and this utility measure based on the rounded frequencies to be published is printed by default (Table 4 gives 0.9456). Hellinger utility based on inner frequencies is also available within the package, though this is of little interest since this table is hidden. Hellinger utility based on expected inner cell frequencies is more relevant. But in any case, a measure based on the published frequencies is arguably more closely related to user needs. Hellinger utility based on cell key perturbed frequencies can be calculated similarly (Table 3 gives 0.9326). This small example data gives no information about the difference between the methods in general. However, for larger datasets it would be fairer to calculate Hellinger utility from the additivity-restored frequencies, since this utility tends to be higher (Table 6 gives 0.9481). Then the risk and utility measures are also calculated from the same data. It is worth noting that new random values within the cell-key method will produce other utility values. In this small example, the Hellinger utility varies a lot. For cell-key perturbed and the corresponding additivity-restored frequencies 95% intervals are [0.8871, 0.9392] and [0.8920, 0.9659], respectively. There is also some randomness within the rounding algorithm. In this case, there are only two equally probable utility values, 0.9457 and 0.9460.

6 Risk

When comparing protection methods, how to measure risk is a crucial discussion. Some cases that can be considered unacceptable are:

- If the existence of certain non-zero frequencies can be concluded with certainty.
- If a large proportion of the perturbed frequencies (especially the small ones)

are equal to the original frequencies.

- If a sensitive variable (party) can be disclosed from quasi-identifiers (age and sex)

Risk measures for all of these aspects may be based on the perturbed data directly (Tables 3 and 4). In the case of CKM, the additivity-restored data (Table 6) can be an improvement, if the decimal values are treated appropriately. Below we will discuss how the estimated inner cell frequencies (Tables 7 and 8) can be used as a basis for risk assessments. Tables 6 and 4 are the corresponding aggregated versions.

A data set of inner frequencies is a compact way of storing a micro data set. Many rows with the same record are replaced with a single row and a frequency value. When the frequencies are perturbed, the data can be viewed as a form of synthetic micro data. We will look at inner frequencies in this way, even when they are not whole numbers. When discussing synthetic data, direct disclosure from quasi identifiers is often important. From the data, one can calculate the probability of guessing the correct political party for all combinations of age and sex (the most frequent was guessed). Such probabilities calculated from original data and the two types of perturbed data are given in Table 9. Only ones (exact disclosure) or probabilities close to one are problematic. One can argue that the frequencies are important in addition to the probabilities, and this is common when looking at synthetic data. However, it is debatable whether it really is twice as bad to reveal two similar units as to reveal a unique unit. One possibility may actually be to treat all probabilities equally. In this example, exact disclosure is found in two of the cell-key cases. Only one of the cases corresponds to exact disclosure in the original data.

Most people know enough about themselves to be able to place themselves in a table to which they contribute. This knowledge can be utilized in order to try to disclose information about other units by removing oneself from the table. This base assumption is a special case of the direct disclosure approach discussed in Lupp and Langsrud (2021). In the following we discuss how this can be used to comparably estimate risk across different perturbation methods. One possibility is to remove oneself from Tables 6 and 4 and then recalculate Tables 7 and 8. A simpler approach is to remove oneself from Tables 7 and 8 directly. The latter approach has been used to re-calculate the probabilities and the results are presented in Table 10. That is, the probabilities are increased by removing a number, maximum one, from a category. Additional disclosure is found. Now both disclosures for cell-key corresponds to disclosure in the original data.

Targeted correct attribution probability (TCAP) has been introduced as a relevant risk measure for attribute disclosure attacks (Taub et al., 2019). It captures the proportion of records, among those considered revealable from the synthetic data (l-diversity of 1), that have the same target value on it's original

equivalent. We will look at the problem similarly and l-diversity of 1 corresponds to ones in the probability tables.

However, we propose a risk measure that is not based on counting records, but rather based on number of exact disclosures in the probability tables. Let a and b be the number of exact disclosures found from the original and perturbed data, respectively. In addition, let c be the number of exact disclosures that the original and perturbed data have in common, i.e., their intersection. In the calculation of b we omit combinations of quasi-identifiers with no corresponding observations in the original data. This corresponds to the intruder scenario on which TCAP is based. To assess disclosure risk, the ratios c/a and c/b provide valuable insight (the latter of which can be viewed as an alternative to TCAP). Indeed, these ratios correspond to established measures used in the fields of information retrieval and machine learning: $m_r = c/a$ and $m_p = c/b$ are known as *recall* and *precision* respectively (Kent et al., 1955). Both measures provide different views on the effectiveness of the perturbation method. Recall provides a measure of how many actual disclosures the perturbated dataset contains, whereas precision measures how many of the disclosures in the perturbed dataset are real. As such, contrary to their common uses within information retrieval one wishes to minimize precision and recall in the context of disclosure.

In order to provide a single risk measure with which to compare perturbation methods, we propose using the well-established F_β measure to combine precision and recall (Rijsbergen, 1979). In general, F_β provides a weighted harmonic means of precision and recall, where β provides a weight for recall, i.e., for $\beta > 1$ recall is prioritized over precision. Thus, our proposed risk measure is as follows:

$$\text{risk} = (1 + \beta^2) \cdot \frac{m_p m_r}{\beta^2 m_p + m_r} = \frac{(1 + \beta^2)c}{\beta^2 a + b} \tag{5}$$

This value lies between 0 and 1, where a value of 0 states that none of the perturbed disclosures are actual disclosures and a value of 1 describes that the disclosures in the perturbed data set are precisely the actual disclosures. Though this is open for debate and yet to be tested extensively, we believe it reasonable to set $\beta < 1$ and place greater weight on precision. As such, in the following we present the calculated risk using the F_β risk measure for $\beta = 0.5$.

Based on Table 9, the SCR risk is 0.0000 (a=1, b=c=0) and the CKM risk is 0.5556 (a=c=1, b=2). Based on Table 10, the SCR risk is 0.7143 (a=3, b=c=1) and the CKM risk is 0.9091 (a=3, b=c=2). New random values within the cell-key method will produce other risk values. From regenerated versions of Table 9, the risk is 0.0000 and 0.5556 in 49% and 48% of the times, respectively. Similarly, for Table 10, the risk is 0.0000, 0.7143, 0.9091 and 1.0000 in 36%, 8%, 51% and 4% of the times. In both cases other risk values can also occur. Within the rounding algorithm, according to Table 9, there are only two equally probable risk values,

Table 9: Calculated probabilities of disclosing party by guessing. For cells marked with *, the wrong party is guessed as most frequent (not same as in original).

	young		middle		old	
Source	male	female	male	female	male	female
Original	1.0000	0.7500	0.4500	0.4000	0.5714	0.8750
CKM	1.0000	1.0000	0.5381*	0.4655	0.4878*	0.7187
SCR	0.7326	0.5584	0.4875*	0.4696	0.5334*	0.7715

Table 10: Calculated probabilities, given that you know yourself, of disclosing party by guessing. For cells marked with *, the wrong party is guessed as most frequent (not same as in original).

	young		middle		old	
Source	male	female	male	female	male	female
Original	1.0000	1.0000	0.4737	0.4286	0.6667	1.0000
CKM	1.0000	1.0000	0.5664*	0.4950	0.5660*	0.8295
SCR	1.0000	0.7001	0.5172*	0.4978	0.6146*	0.9134

0.0000 and 0.5556. Correspondingly, for Table 10, the values are 0.7143 and 0.9091. This small example is only meant as illustration of the methodology. We cannot conclude anything general from these results.

7 R implementations

Fixing additivity by least squares can be done by the function `LSfitNonNeg` in the package `SSBtools` (Langsrud and Lupp, 2021b). The function is made to handle large problem instances by using a sparse matrix methodology. This function is an improvement on functions for similar problems discussed in Langsrud (2019), which mentinos that the R-package `glmnet` could be used. However, after more thorough study, the glmnet approach cannot be recommended generall due to problems with machine precision. In Langsrud (2019) expected inner frequencies were calculated by `glm` using the Poisson family. This function cannot handle large problems. Now a function for iterative proportional fitting, named `Mipf`, is included in package `SSBtools`, where large sparse problems are handled. Recently, the package `SmallCountRounding` has been extended by the function `PLSroundingFits`, which calculates expected inner cell frequencies (using `Mipf`).

The package `SSBcellKey` (Langsrud and Lupp, 2021a) includes a function for CKM with an interface similar to that of `SmallCountRounding`. This package depends on `ptable` (Enderle, 2021). `SSBcellKey` also includes a function, called `PLSroundingFits`, which calculates additivity-restored fits and expected inner cell frequencies (`LSfitNonNeg` and `Mipf` are used). To estimate inner cell frequencies correctly it is important that empty cells (zero frequency) missing in input data are included in the fitting process. The functions `CellKeyFits` and `PLSroundingFits`

include functionality to handle this problem by adding zero frequency rows (the function `Extend0` in the package `SSBtools` is used). As mentioned above, Hellinger utility is included in the `SmallCountRounding` package, but currently this is not included in `SSBcellKey`. The risk measures included above are intended as a basis for discussion and are not implemented in any R-package.

8 Discussion

In general, when attempting to compare two tabular perturbing methods one could first decide the acceptable level of risk. Thereafter each method is tuned to this risk level. The winning method can then be chosen as the one with highest utility. In practice this is not easy; the most challenging part is how to measure risk and how to decide what risk levels are acceptable. This paper does not attempt to provide complete answers to these questions. Rather, we discuss some tools and approaches that can be useful. We propose how data can be made comparable as a starting point for comparable risk measures. A specific risk measure has also been presented and computed. This measure is, however, just one of many possibilities. To handle the problem of assessing risk, a single number may not be the best solution.

Using expected inner cell frequencies, as discussed above, is a very general approach. Such data can be viewed as a kind of synthetic data. An advantage of this approach is that different perturbation methods can be treated in the same way and thus compared fairly. Specific parameter settings within the perturbation methods are not taken into account. This is in accordance with the rule that parameter settings should be hidden from the user. Thus, we can assume that this information is also hidden from an attacker. Then risk can also be compared across different choices of tables to be published.

Whether it is appropriate to look at the data in this way will depend on the type of data that is published. The census 2021 data to be published on grids involve several variables that are not crossed. In this case it seems more appropriate to calculate risk more directly from the published or the additivity-restored data. With such data as a starting point, the framework based on disclosure probabilities can still be used and there will be an even closer relationship to the direct disclosure approach to suppression (Lupp and Langsrud, 2021).

References

Enderle, T. (2021). *ptable: Generation of Perturbation Tables*. R package on `https://github.com/sdcTools/ptable` version 0.3.4.

Kent, A., Berry, M. M., Luehrs, F. U., and Perry, J. W. (1955). Machine literature

searching viii. operational criteria for designing information retrieval systems. *American Documentation*, 6:93–101.

Langsrud, Ø. (2019). Releasable inner cell frequencies by post-processing protected tabular data. Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, The Hague, Netherlands.

Langsrud, Ø. and Heldal, J. (2018). An Algorithm for Small Count Rounding of Tabular Data. Privacy in statistical databases, Valencia, Spain.

Langsrud, Ø. and Heldal, J. (2021). *SmallCountRounding: Small Count Rounding of Tabular Data*. R package on CRAN version 0.9.0.

Langsrud, Ø. and Lupp, D. (2021a). *SSBcellKey: Cell-Key Method for Tabular Data*. R package on `https://github.com/statisticsnorway/SSBcellKey` version 0.0.2.

Langsrud, Ø. and Lupp, D. (2021b). *SSBtools: Statistics Norway's Miscellaneous Tools*. R package on CRAN version 1.2.2.

Lupp, D. and Langsrud, Ø. (2021). Suppression of Directly Disclosive Cells in Frequency Tables. Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, 1–3 December 2021, hosted by Statistics Poland.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, 2nd edition.

Shlomo, N., Antal, L., and Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators. *Journal of Official Statistics*, 31(2):305–324.

Taub, J., Elliot, M., Raab, G., Charest, A.-S., Chen, C., O'Keefe, C. M., Nixon, M. P., Snoke, J., and Slavkovic, A. (2019). Creating the best Risk-Utility Profile: The Synthetic Data Challenge. A version: Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality.

Thompson, G., Broadfoot, S., and Elazar, D. (2013). Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics. Joint UNECE/Eurostat Work Session on Statistical Data.