Data Research Access & Governance Network

Statistical disclosure control for HESA Part 1: Review of SDC theory

Contents

1 Basics of SDC for (official) statistical tables

Data collected provides essential statistical information to policy makers, researchers and the general public. However, the release of statistical information may also have an undesirable effect, especially if information is based on very small specific populations or individual entities instead of on sufficiently large groups of individuals. This can result in the calculation and disclosure of an individual.

Statistical Disclosure Control (SDC) is the term for a range of methods and techniques used to protect statistical data in such a way that they can be released without disclosing confidential information. SDC seeks to optimise the trade-off between the disclosure risk and the utility of the protected released data.

1.1 Purpose

On the most direct basis, the goals of SDC are to

- prevent confidential info being released
- prevent confidential info being perceived to have been released

These are easily met by releasing nothing. However, SDC must be seen in the context of the wider goals of the statistics producer; namely making sure that the most value is gained from its statistical investments. Accordingly, good SDC should also aim to

- maximise provision of statistical data for the public good
- be consistent across individual, environments, and outputs
- be applied reliably and efficiently

1.2 Accidental versus deliberate disclosure

A disclosure occurs when (i) either SDC rules are not applied appropriately, or the SC rules are applied but do not prevent a successful attack, and (ii) a successful attack on the statistic is made. Such an attack may be accidental or deliberate. An accidental attack could occur where a user of the statistic spontaneously realises that some confidential information has been released (or example, the maximum salary for a group of workers is shown). A deliberate attack occurs when an ‘intruder’ actively seeks to unpick SDC protections to gain some information (for example, an estranged parent trying to calculate the university where their child is studying).

Whilst deliberate disclosure is less common than accidental disclosure, and generally much harder in practice than in theoretical attack models, it usually forms the basis of SDC protection assessments. This is because (i) protecting against deliberate disclosure is likely to include spontaneous recognition through error as a special case (ii) the consequences of a targeted attack being successful are likely to be much more severe.

For the purposes of this project, we assume that rules have been correctly applied, but that the protected statistic faces a deliberate attack by a knowledgeable attacker. The specific knowledge held by an attacker needs to be considered to create attack scenarios for particular datasets.

1.3 Types of solutions

Broadly speaking there are six approaches to dealing with the risk posed by a statistic:

1. Not publishing (source data reduction)
2. Limiting output
3. Hiding the problems (suppression)
4. Changing the data (noise, rounding)
5. Table re-design
6. Changing the statistic
7. Replacing real data with synthetic data

The last two are not relevant to this project. As we will be considering only frequency tables (counts of observations in categories) and simple magnitudes (means, medians), transformation of the data is not relevant. As HESA statistics are used for operational purposes by universities, synthetic data runs the risk of operational responses based on false and unrepresentative data.

The first is also less relevant, as this is more appropriate to decisions about which variable to release. In this project, we assume that all the variable under consideration have user value. Accordingly this report focuses on (2) to (5).

Option (3) is *perturbatory*; that is, the values being displayed after SDC no longer represent the true values. They *retain precision but lose accuracy*. Options (2), (4) and (5) are *non-perturbatory*; they protect the data by removing detail, and so *retain accuracy but lose precision*.

In addition, it is possible to consider *input SDC* methods to reduce the sensitivity of the data.

2 KEY CONCEPTS AND CROSS-CUTTING ISSUES

In this section we consider conceptual issues which need to be considered when choosing an SDC strategy.

2.1 CHOOSING A THRESHOLD

When checking tables for disclosure risk, the cell threshold (the minimum number of observations in each cell) is the crucial statistic. A threshold rule is applied to linear tabulations to prevent (a) direct re-identification of an individual and confidential data associated with them, and (b) indirect re-identification through differencing. In rules-based environments such as automatic table generation, this is a hard limit on what can or can’t be published. This threshold has two competing goals: to balance usability and confidentiality of outputs. So what is an appropriate threshold?

A single observation in a cell means that the characteristics of the cell respondent are unique and may be associated with confidential information published using the same classification data. Two observations does not allow the general reader to uncover data about either respondent, but it affords each cell respondent an opportunity to find out something about the other (on the assumption that the respondents knows his or her own tabulated values). Three observations guarantees no confidentiality breach, on the assumption that respondents do not co-operate in the re-identification of others. Hence, most standard textbooks use three as the threshold for exposition, as it seems to solve the problem of direct identification.

However, there are two problems. First, it does not solve the problem of differencing (see next section). A higher threshold may provide better protection, but at a cost in terms of the range of statistics which could be produced. Second, applying the threshold to cell counts may not appropriate when the data are multi-level; for example if a table cell lists twenty students but they all come from the same university, there may be a breach of confidentiality for the university.

As a result, many statisticians would argue that this leaves little margin for error, and encourages the idea that there is a statistically 'right' answer. Ten is a popular number for both national statistics institute (NSI) and research outputs, but five comes close behind. Unfortunately, there is no statistical justification for any number higher than three[1]; thresholds of 5, 10 or 20 (all used by UK government organisations) reflect organisational preferences rather than objective risk measures

2.2 CLASS DISCLOSURE

Class disclosure occurs when information about the range of values for particular groups are disclosed this then provides insight into characteristics of that population. A classic example is where all or none of the observations come in one category.

2.3 Differencing

This is the biggest problem in SDC, and it has no solution.

Differencing involves an intruder using two or more overlapping tables and subtraction to gather additional information about the differences between them. A disclosure by differencing occurs when this comparison of two or more tables enables a small cell (0, 1 or 2) to be calculated.

Secondary disclosure is an unsolvable problem from a theoretical perspective[2] as all possible past and future outputs are in scope for differencing. Even if secondary disclosure checks are restricted to an archive of N previous outputs, there remains a major computation problem. It would be possible to check a new output against the archive items individually in linear time – that is to say, that N pairwise checks would be needed. If it was felt that the new output would be compared against all possible differenced pairs of existing outputs, there would be O(N2) of these; comparing against triples of outputs would scale as N3, etc. By contrast, Moore's Law suggests that computing power doubles every 18 months. In other words, the number of potential comparisons easily outstrips the computing power needed to make those comparisons.

The pragmatic approach followed by most organisations is that a higher threshold reduces the *likelihood* of disclosure by differencing. If the threshold is three, it is quite likely that table cells of 3, 4 or 5 may occur, allowing differences of one or two observations to be revealed. If on the hand, the minimum cell count is 10,000, it seems unlikely that table cells differencing by one or two observations would be routinely generated. So, somewhere between 3 and 10,000 there is a sensible threshold that reduce risk substantially, while allowing most useful outputs to be produced. Unfortunately, there is no agreement on this[3].

2.4 CONSISTENCY AND LINKED TABLES

Consistency of SDC approaches matters because different controls standards applied to the same data increases the chance of disclosure by differencing; for example, rounding to three in one table and round to five in another. Best practice is to apply the same rules to all statistics as far as is

[1] Ritchie, F. (2019, October). 10 is the safest number that there's ever been. Paper presented at Workshop on statistical data confidentiality 2019, The Hague

[2] Ritchie, 2019, above.

[3] Ritchie, 2019, above

reasonable; if necessary, to have two or three ‘classes’ of data to which different rules can be applied; but never to apply different rules on the basis of the output.

A similar problem is created by producing the same information through separate processes: for example, producing frequencies and applying SDC, and then separately calculating frequency proportions from the source data and applying SDC separately. Best practice is to ensure that, as far as possible, tables are derived from each other. In the above example, this would mean calculating the proportions from the (disclosure-controlled) frequency table, not the source data.

2.5 STRUCTURAL ZEROS

A structural zero is where a zero would be expected eg ‘no 17-year olds enrolled at UWE in 2021[4]’, as it is quite reasonable that only 18-year olds post-A Level would apply. In contrast ‘no 17-year olds enrolled at Stirling University in 2021’ is a non-structural zero, as Scottish students can apply to university after their Highers at age 17. The second statement therefore conveys information about enrolment at Stirling which the first did not about enrolment at UWE.

Deciding whether a zero is structural or not requires information on the context, and so is not easily dealt with in an automatic solution. In practice, SDC solutions tend to treat all zeroes as non-structural on the basis that, if the zero is genuinely structural, ‘this cell is zero, as expected’ and ‘this cell is suppressed due to very small numbers, as expected’ are equivalent.

2.6 TABULAR VERSUS PRE-TABULAR METHODS

Tabular SDC techniques are those that are applied to the output tables without the need to consult the source data. This makes them simple to implement, and may be clearer to users. Tabulation methods include suppression, rounding, noise addition and re-categorisation.

Pre-tabular techniques are ones that must have access to the original microdata. A high level of perturbation may be required in order to disguise all unsafe cells. Pre-tabular methods have the potential to distort distributions in the data, but the actual impact of this will depend on which method is used and how it is applied. It may be possible to target pre-tabular methods towards particular areas or sensitive variables. Generally pre-tabular methods are not as transparent to users of the frequency tables and there is no clear guidance that can be given in order to make adjustments in their statistical analysis for this type of perturbation. Pre-tabular methods include cell-key adjustment and differential privacy.

2.7 DOMINANCE

Dominance is the idea that one observation could account for most of the value in a statistical measure, and therefore be identifiable. It can sometimes apply to individuals, but is more of a concern for business statistics where firms might dominate a market or sector. Dominance is managed by the ‘p%-rule’: ordering observations 1..N with the largest first, there is no dominance if the sum of observations 3..N is at least p% of largest observation. Dominance can also be managed by an ‘N, K’ rule, where the largest N observations must contribute less than K% of the total.

Dominance is not relevant for frequency tables, which are expected to be the bulk of HESA outputs. However, HESA does publish some magnitude tables (eg statistics on mean tariffs, academic salaries).

[4] All statement such as this in the report are fictional examples.

2.8 Maxima, minima, percentiles and other rank orderings

As well as dominance, magnitudes tables can lead to problems with rank orderings; for example, noting that the highest salary in a university is X, that the lowest tariff on a course is Y, or that lecturers in Alliance universities all earn below the Zth percentile of national earnings.

Rank orderings are typically more difficult to deal with than other magnitude measures such as dominance. In real-life situations, dominance is almost non-existent except when tabulating magnitudes with few observations and a highly skewed distribution (such as mean salaries for the five employees whose names begin with a 'z', one of whom is the vice-chancellor). These cases tend to fall foul of frequency thresholds before dominance becomes relevant. In contrast, salary percentiles can easily raise issues of class disclosure.

3 Evaluation of specific techniques

In this section the following 'original' table will be used.

			a	b	c	d	e
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	38	7	11	5	**61**
2		HNC/HND	4	1	1	0	**6**
3		First degree	730	141	212	94	**1177**
4		Other undergraduate	50	10	15	6	**81**
5		Masters taught	227	44	66	29	**366**
6		Doctorate research	50	10	15	6	**81**
7		Postgraduate Certificate in Education	13	2	4	2	**21**
8		Other postgraduate research	19	4	5	2	**30**
9		Other postgraduate taught	139	27	40	18	**224**
10		**Total**	**1270**	**246**	**369**	**162**	**2047**

The threshold for disclosure is assumed to be three. We refer to row and column numbers ie 'd1' shows that NI has 5 students on foundation degrees.

3.1 Limiting output

A way to guarantee non-disclosiveness of output is to limit the range of output, and explicitly check the tables for all potential disclosure risks before releasing any. This is managing by identifying 'hypercubes', fixed combination of variables (eg 2 gender categories x 4 age categories x 3 levels of study) which have sufficient observations to prevent disclosure risk. This works well when there are a small number of categorical variables, with a small number of useful categories in each, as the entire set of outputs can be identified, checked and released. Not all outputs need to be released (for example to prevent a profusion of table choices); the data holder can hold some tables 'in reserve', knowing that if they are needed the necessary SDC checks have already been done.

Statistically, the problem with this approach is that the difficulty of proving non-disclosiveness (via differencing) increases exponentially with the number of potential categories and values. This solution has been popular for Census data: the limited number of categories makes this feasible, and the very large number of observations means that risks remain low even if not every possible combination of tables has been checked for disclosure risk.

From the user perspective, the main problem is that this does not allow for additional table combinations, by design. This can limit specialist use, such as research on under-represented groups.

3.2 Hiding data - suppression

3.2.1 How it works

Cell suppression is a non-perturbative method of disclosure control. Primary suppression means removing cells which fall below the threshold, replacing with “n/a”, “-“, “<3” or something similar. To ensure these cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression.

Primary suppression on the original table would show

			a	b	c	d	e
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	38	7	11	5	**61**
2		HNC/HND	4	--	--	--	**6**
3		First degree	730	141	212	94	**1177**
4		Other undergraduate	50	10	15	6	**81**
5		Masters taught	227	44	66	29	**366**
6		Doctorate research	50	10	15	6	**81**
7		Postgraduate Certificate in Education	13	--	4	--	**21**
8		Other postgraduate research	19	4	5	--	**30**
9		Other postgraduate taught	139	27	40	18	**224**
10		**Total**	**1270**	**246**	**369**	**162**	**2047**

This is clearly unsatisfactory: the values in c2 and d8 can be easily recovered from the row and column totals, respectively. The two solutions to this are (1) adjusting the totals to reflect only the values displayed (2) secondary suppression.

Adjusting the totals is recommended[5] for analysts in research environments generating their own tables. It is easy, and less prone to error, particularly for researchers not trained in SDC. However, it can produce inconsistencies between tables, which can lead to disclosure by differencing. Hence, guidance for official statistics usually suggests secondary suppression.

On the table above, successful secondary suppression could involve replacing c8 with a blank:

[5] ONS (2019) *Safe Researcher Training: canonical slide pack*. September

			a	b	c	d	E
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	38	7	11	5	**61**
2		HNC/HND	4	--	--	--	**6**
3		First degree	730	141	212	94	**1177**
4		Other undergraduate	50	10	15	6	**81**
5		Masters taught	227	44	66	29	**366**
6		Doctorate research	50	10	15	6	**81**
7		Postgraduate Certificate in Education	13	--	4	--	**21**
8		Other postgraduate research	19	4	--	--	**30**
9		Other postgraduate taught	139	27	40	18	**224**
10		**Total**	**1270**	**246**	**369**	**162**	**2047**

This protects both c2 and d8 from being reconstructed and protects the table. Note that there are multiple options; for example blanking b8 and c7, or a2 and a8, would also have worked. Different rules can be applied to work out which cells are best. For example, programs such as tau-Argus (see below) allow a 'cost' of suppression to be attached to each cell, so that, if a cell is highly important in another table, the programme will avoid supressing that cell if it can. Cost-based approaches are useful when frequency tables are supporting magnitude tables (for example, one would want to avoid suppressing a cell with just three oil companies if they account for 95% of industry revenue). It is less relevant when only frequency tables are being produced.

3.2.2 Advantages

Cell suppression cannot be unpicked provided secondary cell suppression is adequate and the same cells in any linked tables are also suppressed. Non-structural zeros can be handled in the same way as other undesirable values. The application of primary suppression is easy to automate. The suppression is obvious to the reader.

3.2.3 Disadvantages

Information loss can be high if more than a few suppressions are required. Secondary suppression removes cell values which are not necessarily a disclosure risk, in order to protect other cells which are a risk. Disclosive zeros need to be suppressed. This method does not protect against disclosure by differencing; indeed it is particularly susceptible to it. This can be a serious problem if more than one table is produced from the same data source (e.g. flexible table generation). When disseminating a large number of tables, it is much harder to ensure the consistency of suppressed cells, and care must be taken to ensure that the same cells in linked tables are always suppressed

3.2.4 Practical implementation

Primary suppression can be applied easily manually or automatically. Optimal secondary suppression can be calculated automatically as long as cross-table cost factors are not involved.

3.2.5 Summary

Easy; visible; unambiguous; very susceptible to differencing.

3.3 Changing data (1): rounding

3.3.1 How it works

Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell. In *conventional rounding*, each cell is rounded to the nearest multiple of the base. The marginal totals and table totals are rounded independently from the internal cells. See below for original table conventionally rounded to five.

			a	b	c	d	E
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	40	5	10	5	**60**
2		HNC/HND	5	0	0	0	**5**
3		First degree	730	140	210	95	**1175**
4		Other undergraduate	50	10	15	5	**80**
5		Masters taught	225	45	65	30	**365**
6		Doctorate research	50	10	15	5	**80**
7		Postgraduate Certificate in Education	15	0	5	0	**20**
8		Other postgraduate research	20	5	5	0	**30**
9		Other postgraduate taught	140	25	40	20	**225**
10		**Total**	**1270**	**245**	**370**	**160**	**2045**

Rounding is considered to be an effective method for protecting frequency tables, especially when there are many tables produced from one data set. It provides protection to small frequencies and zero values (e.g. empty cells), as it is no longer clear whether a zero is genuine or not. The method is simple to implement, and for the user it is easy to understand as the data is visibly perturbed.

A larger base value provides more protection (uncertainty) around the data, but also increases the perturbation. Rounding has more of an impact on small values.

A problem with conventional rounding is that totals and cells may not match up; in the case above, column ‘b’ sums to 240 units, whereas the rounded total is 245. This in theory creates an opportunity for the rounding to be unpicked, but in practice this is not at all easy. Similarly, having many genuine zeros can creates theoretical opportunities but again there is little evidence for this in practice. A lesser problem is that user may spot the inconsistencies and the statistical producer may lose credibility. An alternative is that the row and column totals are calculated from the rounded values; as noted in the previous section; but this can then lead to inconsistency between tables, which in itself creates differencing risks.

Random rounding adds extra uncertainty by shifting each cell to one of the two nearest base values in a random manner. Each cell value is rounded independently of other cells, and has a greater probability of being rounded to the nearest multiple of the rounding base. For example, with a base of 5, cell values of 6, 7, 8 or 9 could be rounded to either 5 or 10. Marginal totals are typically rounded separately from the internal cells of the table (i.e. they are not created by adding rounding cell counts) and this means tables are not necessarily additive.

A variation offering greater security is random rounding to a base which may not be the closest, again based on probability. So, if the base is 5 a number 16-19 is likely to be rounded to 15 or 20, but could, with a lower probability be rounded to 10 or 25, or even further.

Controlled rounding uses linear programming techniques to round cell values up or down by small amounts. Its strength over other methods is that additivity is maintained in the rounded table, (i.e. it ensures that the rounded values add up to the rounded totals and sub-totals shown in the table). This property not only permits the release of realistic tables which are as close as possible to the original table, but it also makes it impossible to reduce the protection by 'unpicking' the original values by exploiting the differences in the sums of the rounded values. However this is a computationally complex operation (for example, tau-Argus could not handle this in its base configuration, and needed an additional computation engine specifically for this task). Controlled rounding can achieve specified levels of protection. In other words, the user can specify the degree of ambiguity added to the cells, for example, they may not want a rounded value within 10% of the true value.

3.3.2 Advantages

Rounding is conceptually simple, visible to users and, despite theoretical concerns, seems to offer a high degree of practical protection. It is less vulnerable to differencing than cell suppression.

3.3.3 Disadvantages

Care must be taken when combining rounded tables to create user-defined areas. Cells can be significantly altered by the rounding process and aggregation compounds these rounding differences. The level of association between variables is affected by rounding, and the variance of the cell counts is increased. Small values are disproportionately affected, particularly if the base value is large.

Neither random nor conventional rounding guarantee that the protection cannot be 'unpicked' and so, in theory, the tables should still be checked for disclosure risk.

3.3.4 Practical implementation

Conventional rounding is simple to implement. Random rounding requires a table generator that can make probabilistic assignments. Controlled rounding requires complex software. Rounding can be applied on a table-by-table basis, and does not require consulting other tables.

3.3.5 Summary

Straightforward, visible, effective in practice, limited theoretical guarantees but less susceptible to differencing attacks.

3.4 Changing data (2): Noise addition

3.4.1 How it works

Tables are adjusted by adding a small amount of random noise so that the true value in the cell is uncertain; in the table below values from -2 to +2 have been added to each cell independently:

			a	b	c	d	E
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	37	8	12	6	**61**
2		HNC/HND	2	3	0	0	**6**
3		First degree	729	141	213	96	**1178**
4		Other undergraduate	50	9	15	8	**81**
5		Masters taught	227	44	66	28	**367**
6		Doctorate research	51	11	15	7	**81**
7		Postgraduate Certificate in Education	12	1	2	4	**21**
8		Other postgraduate research	18	2	3	3	**28**
9		Other postgraduate taught	140	25	40	19	**226**
10		**Total**	**1268**	**244**	**369**	**160**	**2046**

The pros and cons of noise addition are similar to rounding, including the choice of whether to adjust row and column totals independently, or whether the make them reflect the data. Like rounding, in theory noise cannot guarantee protection because the same cell in different tables might have different amounts of noise added, and so there is the possibility of disclosure by differencing; the practical possibility of this seems rather less but needs to be assessed on a case-by-case basis. Disclosure by differencing is more likely than for rounded data: as the noise is added independently, there is more scope for inconsistencies to be exploited (whereas in conventional rounding, for example, a 6 in a base 5 scheme will always be rounded to 5).

Because adding noise to the data disproportionately affects small values, an alternative is to have *multiplicative* noise; that is instead of the cell value C becoming C+N where N is an amount of noise, C now becomes C.N. This may be a more appropriate method where the relationship between cells is important, but it can lead to very large amounts of noise being added if there is a large variation in cell values; and multiplicative noise cannot deal with zero cells.

Again, it cannot guarantee protection against unpicking by differencing across tables; and if the noise is generated on the fly as the table is generated, then it is possible that repeated requests for the same table will allow the true values to be uncovered.

3.4.2 Advantages

Noise addition is conceptually simple, easy to implement, and seems to offer a high degree of practical protection. It is less vulnerable to differencing than cell suppression.

3.4.3 Disadvantages

Cells can be significantly altered, particularly if multiplicative rounding is applied. The level of association between variables is affected by rounding, and the variance of the cell counts is increased. Small values are disproportionately affected if additive noise is used.

The protected table is likely to be less same than a rounded table, and ideally should be checked for disclosure risk. If the table is generated dynamically, repeated requests for the same table generate a differencing risk.

Most importantly, noise addition is not visible to the user, and so can cause confusion/reduce the credibility of the statistics producer if inconsistencies are not explained.

3.4.4 Practical implementation

Additive or multiplicative noise is easy to implement on a table-by-table basis, and does not require consulting other tables.

3.4.5 Summary

Straightforward and effective in practice, but doesn't' guarantee privacy and the lack of visibility may cause credibility problems.

3.5 Changing data (3): Cell-key adjustment

3.5.1 How it works

Cell-key adjustment (CKA) is a noise-addition method, However, unlike traditional approaches CKA adds noise consistently across tables. A noise parameter is randomly assigned to every individual microdata record. When records are combined in cells, a deterministic function is applied to the combined noise values so that the same combination of cells always generate the same noise. As a result, there is no risk from either repeated requests for the same dynamic table, or differencing across tables. This therefore works particularly well when the same information is being presented in different ways, providing protection for flexible tables and safely perturbing large high dimensional hierarchical tables.

In the early versions of CKA, a second-stage was used to re-adjust cell values to column and row totals so that there was no inconsistency when totals were preserved. This is no longer seen as best practice, as the table-specific re-adjustment removes the consistency of noise addition and re-introduces disclosure by differencing. Current good practice is to generate totals reflecting the table cells and acknowledge the non-additivity of tables.

The method may also require careful specification of look-up tables for different types of data or output, particularly if the resulting tables are sparse.

3.5.2 Advantages

CKA provides a much stronger guarantee of security by removing the scope for differencing between tables, and for repeated requests for the same dynamically generated table. It can be applied automatically.

3.5.3 Disadvantages

CKA is more complicated to explain and to implement, compared to regular noise addition. The non-additivity of tables may confuse readers, particularly as the method is less transparent than others and harder to explain in methodological notes. Sparse tables (highly skewed variable distributions) may make the specification of the noise lookup tricky.

3.5.4 Practical implementation

This can be applied automatically and dynamically. Once the noise lookup tables have been specified, implementation is straightforward. However, the lookup tables may require some specialist input at the design stage.

3.5.5 Summary

Secure and flexible but complicated to design, and may confuse readers.

3.6 Changing data (4): Differential privacy

3.6.1 How it works

Differential privacy (DP) is a method of noise addition. It seeks to prevent disclosure by considering what values *could have* been in the dataset, not what actually were, and then adding noise. In most of the literature on differential privacy, the noise added for perturbation is taken to follow a Laplace distribution. The "noise"-the random value that is added - ensures that no single person's inclusion or exclusion from the database can significantly affect the results of queries. This gives a mathematical guarantee of the *probability* of privacy (not, as is sometimes reported, a guarantee of privacy itself), with the allowable probability (the 'privacy parameter') set subjectively by the data holder.

Global DP protects against all possible databases eg a differentially-private estimate of the mean salary in a room takes into account that Bill Gates might enter or into the room. As this can lead to absurd or unhelpful estimates, locally-private DP estimates takes a more restrictive view of what is possible, although this does, in theory, weaken the mathematical guarantee of privacy.

DP is often used to model phenomena with heavy tails or when data has a higher peak than the normal distribution. However, it can generate seriously misleading or unhelpful estimates when reporting on rare events, or when summarising continuous variables with extreme values.

DP has become extremely popular in the private sector because of that notional guarantee of protection. Unfortunately this only holds in the case of a single query to the database. As for standard noise-additional models, repeated queries generate a differencing risk. DP implementations are therefore often associated with limits on queries that can be submitted.

3.6.2 Advantages

In a single query, the nominal guarantee of privacy holds, up to a level of probability.

3.6.3 Disadvantages

Absurd results can be generated with extreme distributions or rare events. Multiple queries create a disclosure risk. Setting the level of acceptable risk (the privacy parameter) is a subjective decision, and can be set at a level which exposes the data.

3.6.4 Practical implementation

The theoretical basis is well-known, and many organisations offer table servers with DP solutions.

3.6.5 Summary

Nominally appealing but inappropriate except for limited queries on well-distributed, non-sparse datasets.

3.7 Table re-design

3.7.1 How it works

Re-design involves changing the categories used to display the data: for example

			a	b	c	d	e
			Country				
			E	W	S	NI	**Total**
1	Level of qualification	Foundation degree	38	7	11	5	**61**
2		First degree	730	141	212	94	**1177**
3		HNC/HND/other undergraduate	54	11	16	6	**87**
4		Masters taught	227	44	66	29	**366**
5		PGCE/other PG taught	152	29	44	20	**245**
6		Doctorate/other PG research	69	14	20	8	**111**
7		**Total**	**1270**	**246**	**369**	**162**	**2047**

In this re-design of the original table, no cells fall below the threshold. The information is also completely accurate (numbers of first degree students are correct) but not as precise as before (not clear how many students are PGCE, and how many on other PG courses). Relationships between cells are weakened. Other examples of table re-design would include, for example, the tabulation of all age groups being replaced by six age groups 18-23 and then ‘24+’.

Table re-design requires an evaluation as to how the data is most meaningfully structured. This requires the statistic producer to focus on the meaning of the variables, rather than the circumstances of a particular table. Because this is more likely to lead to the same classifications being used across tables, it is recommended as the first choice for researchers generating their own statistics. This can also be an exercise for official statistics producers: if certain variables always seemed to present disclosure risks, it may be worthwhile considering whether some categories could be usefully combined or adjusted.

Table redesign requires an understanding of the categories, and how they may sensibly be categorised (it is possible to combine categories automatically on the basis of eg correlations with other variables, but this is less likely to lead to satisfactory results). This means that re-design is not amenable to automatic tools.

Table redesign may be unable to address disclosure risks while still keep categories sensible. In such cases, other methods need to be applied. Hence table re-design is typically a decision taken once at the part of the dissemination strategy, and possibly reviewed as disclosure risks appear in tables, but not applied on a regular basis. It is possible to maintain multiple classifications of the same variables (eg “use all ages in these tables but banded ages in those tables”) but this opens the likelihood of disclosure by differencing.

3.7.2 Advantages

A re-designed table retains accurate information. Combined categories may make more sense than the full range of categories. The classification is clear to the user.

3.7.3 Disadvantages

Reclassification of categories requires subject knowledge. Redesign may be insufficient to address problems, and so other methods are needed.

3.7.4 Practical implementation

Ideally, carried out by subject experts prior to the start of dissemination: only the revised categories are used for dissemination, not the original ones. The optimal choice of categories is likely to be a trial-and-error process.

3.7.5 Summary

Conceptually straightforward, semantically appealing and maintaining the information accurately; but not suitable for confidentiality on the fly, especially in automated systems.

3.8 Input SDC

Confidentiality risk can be reduced by limiting the inherent risk in the source data. Examples include:

- Top-coding: limiting magnitudes (such as salaries) to maximum or minimum values so that they are less informative about individuals.
- Noise addition, to make individual magnitudes less certain
- Record swapping: changing characteristics so that over values remain the same but individuals are less easily identified (for example, swapping the ages of two mature students aged 30 and 50)
- Local suppression: removing individual observations that are too noticeable to deal with easily in other ways (for example, post-graduation occupation "astronaut")

These can be combined with output measures to provide a more appropriate risk-utility balance; for example the UK Census 2011 was primarily protected by record-swapping but is now available through cell-key adjusted tables.

An advantage of input SDC is that it provides consistency amongst data holders: by applying the protection to the source data, the distributor can be certain that the same protection has been applied by third parties.

The concern with input SDC is that it is rarely transparent. Whilst top-coding may be visible, the other methods are, by design, invisible to the reader. Moreover, reducing accurate re-identification does not all risks: if, in the case of the mature students discussed above, the swapping lead to mistaken identity, this would not legally be a breach but the reputational damage could be considerable.

The more important concern is the impact on utility. Input SDC is a blunt tool, as it does not differentiate between the uses of the data. Minimal control runs the risk of confidentiality breaches, but high levels of input SDC can substantially reduce data value unnecessarily. For this reason, input SDC is often limited to the most extreme cases, with output SDC being used to manage risk in specific tables.

3.9 Dealing with magnitude tables

Dealing with magnitudes depends on how important outliers and distributions are. If exact values are not important then limiting extreme values can be a very simple way to protect the data. Ideally, this is done once at the data set rather than table specific: for example, top coding salaries to a maximum of £200,000 in the data, rather than removing the top 1% in a particular table (the latter exposes the table to disclosure by differencing). This will affect some other statistics (the mean, for example, but not the median) but using top-coded values as the 'genuine' data does reduce the problem of disclosure by differencing.

As general rule, uncapped maxima and minima are best avoided unless they are structural (eg a minimum tariff of 0, or a statistic which can range between 0% and 100%).

There is no obvious solution to dominance but as it is very rare and easily tested, it should be possibly for table generators to demonstrate that dominance rules are met.

4 Resources

4.1 Online resources

https://research.cbs.nl/casc/CENEXindex.htm

4.2 Automated tools

4.2.1 Tau-Argus

τ-ARGUS (see Hundepool et al. 2011) is a software package which provides tools to protect tables against the risk of statistical disclosure . Controlled rounding is easy to apply in τ-ARGUS and the controlled rounding procedure (CRP) used was developed by JJ Salazar (see Salazar-González et al. 2006). The technique procedure is based on optimisation and produces rounded tables, where the rounded values add up to the rounded totals and sub-totals shown in the table. This means realistic tables are produced whilst ensuring that original values can not be calculated by unpicking values and contrasting the differences in the sums of the rounded values.

Controlled rounding gives sufficient protection to small frequencies and creates uncertainty about the zero values (i.e. empty cells). In general, cell suppression leaves empty cells unmodified.

4.2.2 sdcTable

sdcTable is freeware tool (Meindl, Statistik Austria) the software provides safety rules for; primary suppression: dominance, P% rule, Frequency or threshold rule. The tool be applied to linked tables and incorporates auditing. Available methods for secondary suppression include Hypercube and HiTaS.SdcTable has been successfully deployed by NSIs, and there is no specialist IT requirements required for implementation of the software (although knowledge of R is required). Disadvantages of sdcTable includes not being able to manage unusual table structures. The documentation recommends that results should be checked to ensure that output is safe.

ANNEX: SUMMARY TABLE

Type	Maintains accuracy?	Maintains precision?	Clear to readers?	Single-table protection	Protection against differencing	Ease of implementation	Suitable for table generators?
Suppression	Yes	Yes	Yes	Good	Poor	Very easy	Yes
Rounding	No	No	Yes	Very good	Good	Very easy – very hard	Yes
Simple additive/ multiplicative noise	No	Yes	No	Very good	Some	Easy	Yes
Cell-key adjustment	No	Yes	No	Very good	Excellent	Easy to apply Lookups require specialist knowledge	Yes
Differential privacy	No	No	No	Excellent	Some	Requires specialist skills	Yes
Table re-design	Yes	No	Yes	Excellent	Excellent	Requires subject knowledge	No