Ten is the safest number that there's ever been

Felix Ritchie

University of the West of England (UWE), Bristol

Data Research, Access, and Governance Network Working Paper Series

No. 2021/02

10 is the safest number that there's ever been

Felix Ritchie

University of the West of England, Bristol.

Email Felix.ritchie@uwe.ac.uk

Abstract: When checking frequency and magnitude tables for disclosure risk, the cell threshold (the minimum number of observations in each cell) is the crucial statistic. In rules-based environments, this is a hard limit on what can or can't be published. In principles-based environments, this is less important but has an impact on the operational effectiveness of statistical disclosure control (SDC) processes.

Determining the appropriate threshold is an unsolved problem. Ten is a common threshold value for both national statistics and research outputs, but five or twenty are also popular. Some organisations use multiple thresholds for different data sources.

These higher thresholds are all entirely subjective. Three is the only threshold which has an objective statistical foundation, but most organisations argue that this leaves little margin for error. Unfortunately, there is no equivalent statistical case for any number larger than three: ten is popular because it is popular. This is particularly the case for research environments, where there is no guidance.

This paper provides the first empirical foundation for threshold selection by modelling alternative threshold values on both synthetic data and real datasets. The paper demonstrates that this is a complex question. The trade-off between risk and value is well-known, but we demonstrate that the protection of a higher threshold depends on the risk measure. There is no monotonic relation between a threshold and risk, as higher thresholds can increase disclosure risk in particular scenarios. The blind application of high-threshold rules might mask new risks. There is no unambiguous result, other than the simplistic ones that more observations reduces risk and higher thresholds reduce utility.

Finally, the paper notes that a reconsideration of disclosure checking practices can reduce risk irrespective of the threshold for some risk scenarios.

Acknowledgements

The first draft of this paper was presented at the 2019 Eurostat/UNECE Confidentiality Workshop; I am grateful to participants for a number of useful comments. I am indebted to Alastair Currie at HM Revenue and Customs for pointing out that the paper refers to 'disclosure risk' when it means 'information recovery', which prompted the rethinking of the paper in terms of two types of 'failure'. I would also like to thank Ben Derrick at UWE for commenting on the paper and the logic of the simulation. All errors remain my own. Full details of the results aand code are available at the project website http://www.fivesafes.org/SDC/10s_paper/.

1 Introduction

When checking frequency and magnitude tables for disclosure risk, the cell threshold (the minimum number of observations in each cell) is the crucial statistic. In rules-based environments, this is a hard limit on what can or can't be published. In principles-based environments, this is the default rule which determines how conversations about acceptable outputs will go (see [1], for a description of the difference between rules- and principles-based checking schemes).

This threshold, often the first rule in any statistical disclosure control (SDC) guide, has to do a lot of heavy lifting. In a rules-based world, that one number has to balance usability and confidentiality of outputs. This is an impossible task for a single measure, and it is straightforward to demonstrate how it can fail to achieve either outcome [2]. In ad-hoc or principles-based environments, the actual value is less important, but a poorly-chosen limit can still affect the efficiency of the environment and the credibility of the organisation setting the rules.

The problem is: what is an appropriate threshold? Three is the only value which has an objective statistical basis, but many practitioners would argue that this leaves little margin for error, and encourages the idea that there is a statistically 'safe' answer. Ten is a popular number for both national statistics institutes (NSIs) and research outputs, but five comes close behind. Some organisations use multiple levels eg five for standard outputs, ten for outputs based on more sensitive data. One organisation uses thirty for research output but less for its own statistics.

NSIs offer training to their own staff and to researchers, but rarely admit to the truth: that ten (or five, or twenty) is a subjective choice. I have observed training courses where the trainers try to defend ten as if it has some inherent, magical power. Trainers who try to do this invariably lose the argument, and thus their credibility, because the statistical case is absent. Ten is popular because (a) it is a memorable round number (b) other people use it. In a world of uncertainty, doing what others do can be the easiest and most defensible option.

For a limit above three, the main rationale is that a higher limit reduces the likelihood of disclosure by differencing. In the early 2000s, some simple statistical analysis (now lost) was carried out using randomly generated data by the Virtual Microdata Laboratory (VML) team at the UK Office for National Statistics (ONS). This suggested that the opportunities for disclosure by differencing decrease very rapidly once cell thresholds rise above five or six. In the research environment managed by the VML team, ten therefore seemed a very safe threshold to require of research outputs, and one which was acceptable to researchers.

At that time, the decision to use ten as the threshold by the VML was unusual, and not even common within ONS; the statistical minimum of three was preferred. Some fifteen years later, ten seems the most popular number applied to research outputs worldwide, as well as becoming more common in official statistics.

However, there is still no statistical case for this, and it is an important question. A higher threshold is expected to reduce risk, but is also expected to restrict the publication of useful outputs. Evidence on the balance of risks is useful to statistics producers keen to maximise the value of data.

Unfortunately, the issue is not amenable to analytical review. This paper instead offers empirical analysis of how thresholds affect risk and utility by modelling alternative threshold assumptions on both synthetic data and on a real dataset used by researchers. The aim is not to prove that any particular threshold is 'best' – this is not possible – but to provide supporting evidence for the subjective decisions that NSIs make.

The paper demonstrates the expected conclusion that more observations reduce risk, and higher thresholds reduce the utility of outputs. However, the novel key findings of the paper are that:

- risk is not a monotonic function of the threshold: a higher threshold can lead to higher risk as well as lower risk, and
- the relationship between risk and threshold varies with the type of risk
- non-statistical measures (in particular, clearer guidance on outputs) can reduce risk more effectively than higher thresholds

There is no extant literature on this topic. The next section therefore introduces the conceptual framework. Section 3 describes the approach taken, and section 4 findings. Section 5 concludes.

2 Conceptual review

2.1 Strong versus weak differencing

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.

A single observation in a cell means that the characteristics of the cell respondent are unique and may be unambiguously 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 (eg [3]) use three as the threshold for exposition: it solves the problem of direct identification with a clear, objective statistical justification.

In contrast, indirect identification through differencing (exploiting different numbers of observations across multiple tables to infer single observations) has no theoretical solution. For any table A there exists a second table B such that (A-B) has single observations in it. NSIs invest considerable time and effort to ensure that A and B are not both generated, but this is not a guarantee of protection. Even if B is not published, how can the NSI guarantee that B could not be created by some combination of some other tables C, D, E, F…? A proof that a table cannot be differenced would require knowledge of every other table produced in the past, present and future on that data, which is clearly impossible.

The theoretical impossibility of proving non-differencing is a straw man: no experienced organisation claims that as its target. However, organisations may have what could be described as a 'strong differencing' policy:

> ***Strong differencing****: thresholds, and the choice of related tables to be checked, are chosen to ensure that there is no reasonable chance of differencing between published tables, given the likely set of published tables*

Strong differencing has two implications.

First, tabular data protection is determined by history: the first table to be produced determines which others may be produced. This is a feasible policy for the official statistics produced by NSIs, where the full range of published outputs is typically planned in advance. However, even NSIs cannot review all possible combinations (this is computationally prohibitive in operational circumstances), and there remains a potential risk [4]. This is much more problematic for research outputs, where table production is determined by the interests of individual researchers on an ad-hoc basis.

The second problem is that strong differencing pays no attention to the value of published outputs. While the publication of confidential data is clearly problematic, the non-publication of non-confidential data due to unfounded confidentiality concerns can lead to public benefits being lost.

Strong differencing relies upon the assumption that the ability to uncover a cell value through differencing implies a breach of confidentiality. This is clearly not true. A single observation in a cell may disclose information about the individual; in practice, this is unlikely, except in cases where extreme values are being discussed (for example, the highest earner in a small geographical area).

Avoiding cell counts of one or two to prevent direct identification seems a sensible precaution, as such small cells are also likely to be of little value. There is also an argument that avoiding small numbers is important for the NSI or data holder to publicly demonstrate that it is not taking risks with confidentiality. In contrast, it is not at all clear that the same standards need to be applied to small counts arising from differencing; these require the difference to be noticed, as well as present.

An alternative approach might be described as a 'weak differencing' policy:

> ***Weak differencing****: thresholds, and the choice of related tables to be checked, are chosen to ensure that the likelihood of differenced values being disclosive is balanced with the likely loss to public benefit of not producing the tables.*

This differs from strong differencing by acknowledging three things:

- The reasonable possibility of differencing
- The uncertain disclosiveness of differenced tables
- The potential loss from unrealised public benefit

This is much more explicitly a risk-benefit model, with the risks and benefits being very subjective. As a result, the perspective of the decision-maker has a strong influence over the table-checking regime and the choice of threshold.

For example, the author has encountered 'default-closed' [5] data holders who argue that the public benefit of any particular table in social science research is negligible; hence, the possibility of disclosure by differencing must be exceedingly low to be outweighed by the benefit. In contrast, data holders following the EDRU ethos [6, 7] would assume that the public benefit has already been established by the decision to use the data for research or official statistics, and therefore the onus is on those suggesting a cell be suppressed to prove the substantive case for a breach.

2.2 The choice of threshold

NSIs and other data holders, if they describe any policy on differencing, typically cite a strong differencing model as this allows them to establish credibility in protecting confidentiality. As noted, this is feasible for official statistics. However, for ad hoc and research outputs, most organisations apply weak differencing (even if default closed), and so the choice of threshold is highly subjective.

In 2003 ONS's Virtual Microdata Laboratory (VML), a secure facility for researchers, began using a threshold of ten instead of the three then in use. This was justified by (1) reference to Monte Carlo simulations of differencing (now lost) which showed the likelihood of difference became negligible after a threshold above 5; and (2) an analysis [8] which argued that this gave confidence that simple threshold checks would also deal with the problem of multiple respondents from the same business when dealing with hierarchical data. However, a primary motivation for the specific choice of ten was that it was high enough to avoid questions of differencing but also acceptable to researchers (source: personal discussion).

The VML was not the first such research centre, but since 2003 the number of them has grown steadily, and almost all use a threshold higher than three. Ten appears to be the most popular, but we are not aware of any justification other than that this seems to be popular. In other words, everyone uses ten because everyone else uses it. In a world where data holders face considerable pressure to show that they are not unduly taking risks, following common practice is a sensible strategy.

This is not universal. In just the secure facilities in UK public sector, values from three to thirty are used. One organisation uses five as its default, but raises the threshold to ten for more 'sensitive data'. This, it seems likely, is primarily to demonstrate that some data is more sensitive/risky and that the organisation is taking a more active approach than just applying a blanket rule.

All discussions about confidentiality protection involve a large amount of subjective reasoning [9, 10]. However, for the threshold rule this is complicated by the apparent absence of any objective statistical evidence.

Two approaches may be considered to improve data holders' confidence in their judgments. One is to create tables from a genuine research data source, and evaluate the impact alternative thresholds might have had on both disclosure and usability. The alternative is to carry out the same analysis but using simulated datasets to investigate the effect of different data profiles.

Both of these approaches are tried here. The analyses cannot be definitive, as they are specific to the context (either categories chosen for the real data, or the simulation characteristics). Rather, the aim is to explore whether sufficiently general lessons can be learned from trying a range of alternative specifications.

3 Method

We tackle this issue by considering three cases which seem to present the most obvious problems. We consider two risk measures. The first is that cell counts of 1 and 2 are the values to avoid, irrespective of the formal threshold. The second is that cell counts below the threshold are not uncovered (for example, if the threshold is 5, that the difference between two tables is also not less than 5). The choice of measure has a significant effect on results.

3.1 Case 1: differencing between a set and a subset

In this case we assume a situation as in table 1 and 2:

Table 1 Residents

Age	Urban	Rural	Total
50-54	20	12	32
55-59	23	13	36
60-64	26	14	40
65+	28	14	42
	97	53	150

Table 2 Homeowners

Age	Urban	Rural	Total
50-54	20	11	31
55-59	23	11	34
60-64	26	14	40
65+	27	11	38
	96	47	143

Tables 1 and 2: Example of differencing in subset

There is an implicit table 2a here where many 1s and 2s occur:

Table 2a Non-homeowners

Age	Urban	Rural	Total
50-54	0	1	1
55-59	0	2	2
60-64	0	0	0
65+	1	3	4
	1	6	7

Table 2a: The implicit differenced table

In this example, a threshold of 3 would lead to many 1s and 2s being generated by differencing. A threshold of 35 is necessary to prevent and 1s and 2s arising from differencing (if totals are included), but would lead to results less than the threshold being uncovered; in fact, for these particular tables, there is no threshold that prevents uncovering of values below the threshold.

To consider this option, we:

- Create random category allocation for Age (X)
- Create random u/r category allocation for residents (Y) with p_{urban} > 50%
- Create random home/rent category allocation (Z) with $p_{homeowner}$ > 50%
- Tabulate X:Y and X:(Z=homeowner), correcting for the threshold (zero is deemed below threshold and redacted)
- Tabulate X:(Z=renter) and count (a) number of 1s/2s in cells where the originals were not suppressed (b) number of cells in the differenced table which fall below the threshold
- Store number of 1s/2s, number of uncovered cells, mean observations and median observations of X:Y and X:(Z=renter)
- Iterate N times with new random values

The proportion of 'bad cells' reported below is the proportion in the differenced Table 2a ie a score of 100% would mean that every cell in Table 2a can be recovered, and contains either a 1 or 2 or a value below the threshold, depending on the failure criterion.

3.2 Case 2: Row totals revealing suppressed cells

Consider Table 3, where a threshold of 5 has been applied, placed alongside Table 1 for clarity:

Table 1 Residents				Table 3 Education			
Age	Urban	Rural	Total	Age	No degree	Degree	Total
50-54	20	12	32	50-54	26	6	32
55-59	23	13	36	55-59	29	7	36
60-64	26	14	40	60-64	36	<5	36
65+	28	14	42	65+	39	<5	39
	97	53	150		130	13	143

Tables 1 & 3: Example of differencing through row totals

We assume that tables are presented with values below the threshold suppressed, and totals adjusted to reflect suppressions. This is considered good practice in analytical environments. Officials statistics are more likely to use secondary (within-table) suppression to maintain marginal totals. Neither solution provides protection against cross-table differencing, but we take the former route as (a) this offers more and simpler opportunities for differences to arise, and (b) this can be programmed without the need to define secondary suppression rules.

Although Table 3 has the marginal totals adding up to the displayed values (and so the missing values cannot be reconstructed from this table), it is clear that a comparison of Tables 1 and 3 reveals the suppressed values.

Table 3 is the worst-case scenario: If there were more than two categories in Table 4, then row totals would not necessarily be sufficient to expose suppressed values.

In this case, a threshold of 3 would have avoided this problem as the low values would not have been removed. A threshold of 15 would also have avoided thee problem as the 'rural' column in Table 1would also have been hidden.

To consider this worst case, we

- Use X and Z, as above
- Create random binary category allocation for Qualifications (Q) using p_{degree}% such that one category is relatively rare
- Tabulate X:Z and X:Q, correcting for the threshold and dropping rows in X:Z with no valid values (zero is below threshold)
- Compare row totals
- Store number of exposed cells (in **both tables**), mean observations and median observations of X:Z and X:Q

3.3 Case 3: Direct disclosure by negation

Finally, consider Table 4:

Table 4 Ethnicity

Age	Urban	% white	Rural	% white
50-54	20	90%	12	92%
55-59	23	87%	13	92%
60-64	26	85%	14	79%
65+	28	89%	14	93%
	97	88%	53	89%

Table 4: Example of differencing through complements

As counts of humans must be integers, the complementary Table 4a can easily be determined:

Table 4a Non-white frequency

Age	Urban	Rural
50-54	2	1
55-59	3	1
60-64	4	3
65+	3	1
	12	6

Table 4a: The implicit low-frequency table

In this case, it is likely that only a very high threshold would address this problem; a better guideline might be that, when binary conditions are tabulated, the smaller fraction should always be displayed.

To consider this case, we

- Use X and Z, as above
- Create random binary category allocation for Ethnicity (E) using p_{white}% such that the negative (non-white) is very rare
- Tabulate X:W and X:(1-W), allowing for the threshold checks on the numbers themselves, but not on the percentages (ie X will be tested against the threshold, not whether X*p% is below)
- Record number of implicit 1s and 2s or uncovered cells (we don't test for zero, so assume these are structural for simplicity)
- Don't count the cells where the source number is supressed.

For this, we could just have chosen rural or urban, so why both? The aim is to give a better sense of missed values: as a checker, high initial frequencies (w=urban) are less likely to cause concern, but if the initial frequencies are low (w=rural) this might signal potential problems. Running this way covers both options.

3.4 Generating simulated data

Data were initially generated using the following parameters

- Number of iterations: 1,000
- Number of observations in the dataset: 500, 1,000, 5,000 and 10,000
- Number of X categories: (a) 10 uniformly distributed and (b) 5 dominated by one category
- Values of p% (urban): 70%, 80%, 90%, 95%
- Values of p% (homeowner): 70%, 80%, 90%, 95%
- Values of p% (degree): 15%, 10%, 5%
- Values of p% (white): 90%, 95%, 99%
- Thresholds evaluated: each of 3-15, 20, 25, 30 (16 in total)

Initially various combinations of values were entered. However, because (as will be shown later) the relationship between sample characteristics and risk potential is highly non-linear, the program was recoded to automatically generate and store multiple parameters values for graphing.

The same exercise was then carried out on three genuine datasets, with real variables taking the place of the simulated variables 'urban', 'homeowner', 'degree' and 'white':

	Charity[1]	Teaching LFS[2]	LFS low-paid[3]
Data source	Published accounts	Employee survey	Employee survey
Observations	686	19,032	4,859
X (‘age’)	‘year’: 2010 83 2011 150 2012 151 2013 153 2014 149	‘age’: 50-54 6,590 55-59 6,366 60-64 5,119 65-69 957	‘age’: 50-54 2,091 55-59 1,860 60-64 850 65-69 58
Y (‘urban’)	‘big’: 49%	`female’: 52%	`female’: 58%
Z (‘homeowner’)	‘survivor’: 65%	`england’: 82%	`england’: 84%
Q (‘degree’)	‘secure’: 6%	`degree’: 11%	`degree’: 4%
W (‘white’)	‘surplus’: 96%	‘white’: 97%	‘white’: 98%

[1] Green et al [11]. ‘Survivor’:still trading 2015. ‘Secure’ and ‘surplus’ relate to financial viability

[2] Labour Force Survey Teaching Dataset, UK Data Service dataset SN4736. Gender, ethnicity and age randomly perturbed; employed and age 50+ only

[3] LFS data as above, restricted to subset earning under £10/hour

Table 5: Datasets used

Genuine variables were relabelled as X, Y, Z, Q and W to allow the same code as the simulated data to be run. The same thresholds were evaluated in the true datasets as in the simulated data but without multiple iterations and without different values for the y, z, q, or w percentages.

The code produced, for both simulated and genuine datasets:

- The proportion of ‘bad’ results depending on the measure (that is, either the the number of 1s and 2s uncovered, or the number of cells below the threshold which were exposed
- The proportion of ‘ok’ results (that is, the number of usable cells once thresholds had been applied

Note that %“bad” + %“ok” + %”suppressed but not bad” = 100%. These are stored for every combination of thresholds and (for simulations) values of the simulated characteristics.

The program is written in Stata and can be downloaded from http://www.fivesafes.org/SDC/10s_paper/, along with datasets and full output files.

4 Results

4.1 Simulated data

The simulations produce a very large number of results: 2 types of failure (1-2, uncoverd cells), 2 types of data distributions (uniform/skewed), 16 thresholds, 4 X categories, 3 or 4 other categories, and four sample sizes. This section therefore summarises key features rather than going though in detail. Only the uniform distribution is reported on, and only results for 500 and 5000 observations.

In the tables below, 'bad % (1/2)' is the proportion of the original table cells that generate an uncovered cell of 1 or 2 observations; 'bad % (uncovered)' is the proportion of uncovered cells below the threshold; 'usable %' is the % of cells in the original tables which were not suppressed for being below the threshold. Results are depicted in 10% categories. Annex 1 expands the results below to the full set of thresholds. The log files are available at the above website.

4.1.1 Case 1

Table 6 shows the proportion of cells in the implied Table 2a marked 'bad'as defined by uncovered 1/2s, and the proportion of times (out of 1,000 iterations of 16 cases) that this proportion was observed. For example, with 500 observations and a threshold of 3, in 5% of the 16000 cases no cells with 1s or 2s were uncovered (0% bad); with 5000 observations and a threshold of 30, between 10.1% and 20% of the cells were uncovered in 4% of the 16,000 cases. A blank space means no cases occurred at that 'bad' proportion. A 0% means that less than 0.5% of combinations generated that proportion of errors.

	500 observations				5000 observations			
Bad %	Threshold				Threshold			
(1/2)	3	10	15	30	3	10	15	30
0%	5%	33%	45%	52%	65%	65%	66%	78%
10%	21%	23%	19%	19%	15%	15%	15%	15%
20%	23%	14%	13%	14%	8%	8%	8%	4%
30%	21%	12%	12%	11%	8%	8%	8%	3%
40%	14%	8%	7%	4%	3%	3%	3%	1%
50%	10%	6%	3%	0%	0%	0%	0%	0%
60%	5%	3%	1%					
70%	2%	1%	0%					
80%	0%	0%						
90%								
99%								
100%								

Table 6: Bad cells (1s/2s) in case 1

As expected, a higher number of observations reduces the proportion of 'bad' results (ie where the gap between two non-supressed cells is 1 or 2). A higher threshold also monotonically decreases the ‘bad’ proportion. However, this is not the case if the criterion for ‘bad’ is ‘uncovered cells below the threshold’; see Table 7.

	500 observations				5000 observations			
Bad %	Threshold				Threshold			
(<threshold)	3	10	15	30	3	10	15	30
0%	5%	4%		1%	65%	31%	21%	20%
10%	21%	9%	0%	6%	15%	7%	5%	6%
20%	23%	6%	2%	10%	8%	6%	6%	5%
30%	21%	9%	7%	17%	8%	9%	9%	6%
40%	14%	10%	15%	28%	3%	14%	15%	6%
50%	10%	29%	56%	38%	0%	32%	35%	29%
60%	5%	11%	13%			0%	7%	8%
70%	2%	9%	6%				3%	2%
80%	0%	7%	1%				0%	1%
90%		5%	0%				0%	3%
99%		1%						4%
100%		0%						12%

Table 7: Bad cells (uncovered below threshold) in case 1

While it remains true that more observations reduces the proportion of bad cells (there is always a higher proportion of ‘0% bad’ for 5000 obervations at each threshold), it is longer the case that a higher threshold reduces the number of bad cells; if anything, a higher threshold is more likely to lead to cells below the threshold being uncovered. However, it is also clear that this is highly non-linear in the proportion of bad cells being uncovered. A higher threshold increases the number of suppressed cells, but increases the chances that a differenced cell falls below the limit. This is not immediately amenable to modelling.

As would be expected, the usability of the data depends significantly on the number of observations and the threshold. Table 8 shows the proportion of non-suppressed cells in the source tables (Table 1 and Table 2, excluding totals, for case 1). Thus, with 500 observations, in 49% of the 16,000 cases no cells were suppressed whenthe threshold was 3; but when the threshold was raised to 30, in no cases were more than 50% of the cells in Table 1 or Table 2 unsuppressed.

	500 observations				5000 observations			
	Threshold				Threshold			
Usable % (unsuppressed)	3	10	15	30	3	10	15	30
0%								
10%								
20%				1%				
30%				11%				
40%				24%				
50%	0%	43%	60%	63%				3%
60%	3%	9%	17%					19%
70%	12%	10%	12%					3%
80%	10%	11%	10%				0%	0%
90%	13%	11%	2%				2%	0%
99%	14%	12%	0%			2%	12%	8%
100%	49%	4%			100%	98%	86%	67%

Table 8: Usable (non-suppressed) cells in case 1

The sharp break at 50% is because the value of both the high-density and low density columns are being counted. With a small number of observations, the low–density column is completely suppressed. With a high number of observations, it takes a very high threshold before even the low-density column is suppressed. Again, there is a non-linearity in the usable proportion.

Beyond the obvious point that more observations or lower thresholds reduce the number of suppressions, it is not clear how to model the relationship between observations, threshold and usablity.

4.1.2 Case 2

For this case, there are 12,000 outcomes (1000 iterations by 4Y and 3Q proportions). Table 9 shows the number of bad cells by 1s and 2s.

	500 observations				5000 observations			
Bad %	Threshold				Threshold			
(1/2)	3	10	15	30	3	10	15	30
0%	22%	57%	77%	93%	100%	100%	100%	100%
10%	16%	22%	14%	6%				
20%	12%	9%	5%	1%				
30%	11%	5%	3%	0%				
40%	12%	3%	1%	0%				
50%	13%	2%	0%					
60%	9%	1%	0%					
70%	4%	1%	0%					
80%	2%	0%	0%					
90%	0%	0%						
99%								
100%								

Table 9: Bad cells (1s/2s) in case 2

Again, more observations and a higher threshold recudes the number of problematic cells; with 5000 observations, none of the combination of parameters leads to a 1 or 2 being uncovered by differencing. But, as with Case 1, changing the definition of 'bad' to mean 'cells below the threshold being exposed' gives a very different story; see Table 10.

		500 observations				5000 observations		
Bad %		Threshold				Threshold		
(uncovered)	3	10	15	30	3	10	15	30
0%	22%	11%	33%	42%	100%	100%	94%	50%
10%	16%	15%	25%	32%		0%	6%	1%
20%	12%	13%	12%	17%			0%	2%
30%	11%	9%	6%	7%			0%	2%
40%	12%	8%	5%	2%				2%
50%	13%	8%	6%	0%				2%
60%	9%	8%	7%	0%				3%
70%	4%	7%	5%	0%				8%
80%	2%	8%	1%					12%
90%	0%	8%	0%					13%
99%								
100%		5%	0%					5%

Table 10: Bad cells (uncovered) in case 2

When considering row differences the question of bad cells (where the row totals in Table 1 allow the missing values in Table 3, or vice versa, to be uncovered) is more complex. With 500 number of observations, then a threshold of 10 performs worse than either a threshold of 3 or 30. On the other hand, with 5000 observations, a always performs worse than a lower one. As the number of observations increases, a higher threshold increases the chance that one or other row (but not both) has just one cell suppressed, creating an exploitable difference.

The results on usable cells are also complex; see Table 11.

	500 observations				5000 observations			
Usable %	Threshold				Threshold			
(unsuppressed)	3	10	15	30	3	10	15	30
0%								
10%								
20%								
30%								
40%				1%				
50%		11%	33%	71%				0%
60%	0%	40%	47%	27%				6%
70%	1%	22%	19%	0%				2%
80%	9%	25%	1%					30%
90%	34%	2%						11%
99%	34%	0%				0%	6%	0%
100%	21%				100%	100%	94%	50%

Table 11: Usable (unsuppressed) cells in case 2

Again, more observations and a lower threshold both lead to more cells being unsuppressed; but the non-linearity in the proportion of cells that are usable is more pronounced than in Case 1.

4.1.3 Case 3

As for Case 2, there are 12,000 outcomes (1000 iterations, 4Y and 3W proportions). Table 12 presents results for uncovered 1s/2s.

	500 observations				5000 observations			
Bad %	Threshold				Threshold			
(1/2)	3	10	15	30	3	10	15	30
0%	1%	3%	4%	5%	28%	28%	28%	36%
10%	9%	18%	23%	32%	16%	16%	16%	27%
20%	24%	27%	31%	34%	16%	16%	16%	11%
30%	24%	23%	24%	22%	19%	19%	19%	10%
40%	18%	14%	11%	6%	13%	13%	13%	9%
50%	14%	9%	5%	0%	6%	6%	6%	5%
60%	7%	5%	1%		2%	2%	2%	2%
70%	2%	1%	0%		0%	0%	0%	0%
80%	0%	0%	0%		0%	0%	0%	0%
90%								
99%								
100%								

Table 12: Bad cells (1s/2s) in case 3

When considering the potential for exposure of binary complements, there appears to be an issue, even with a threshold of 3, when the number of observations is small. More interestingly, increasing the number of observations brings results for the lower thresholds very much in line with the higher ones across all thresholds (the differences between threshold are less than 0.5% ie 60 occurrences). This is not something observed in Cases 1 and 2.

When considering unsuppressed cells, the situation worsens; see Table 13.

		500 observations				5000 observations		
Bad %		Threshold				Threshold		
(uncovered)	3	10	15	30	3	10	15	30
0%	1%	0%	0%	1%	28%	4%	0%	1%
10%	9%	6%	7%	9%	16%	5%	1%	5%
20%	24%	16%	16%	16%	16%	4%	4%	2%
30%	24%	10%	9%	10%	19%	7%	6%	0%
40%	18%	6%	8%	18%	13%	13%	10%	1%
50%	14%	31%	41%	45%	6%	35%	34%	23%
60%	7%	9%	9%		2%	7%	16%	16%
70%	2%	9%	6%		0%	7%	11%	8%
80%	0%	7%	4%		0%	8%	9%	9%
90%		5%	1%			7%	8%	11%
99%		1%	0%			2%	2%	7%
100%		0%				1%	1%	16%

Table 13: Bad cells (uncovered) in case 3

As in Cases 1 and 2, using uncovered cells as the measure of failure menas that a higher thresohld is associated with more errors; once more, there is no linear relationship between thresholds in terms of the proportion exposed.

More observations does increase the number of usable cells, but there remains a large information loss associated with the higher threshold; see Table 14.

	500 observations				5000 observations			
Usable %	Threshold				Threshold			
(unsuppressed)	3	10	15	30	3	10	15	30
0%								
10%								
20%								
30%				1%				
40%				13%				
50%	0%	43%	60%	86%				3%
60%	2%	8%	15%					15%
70%	10%	6%	7%					6%
80%	10%	12%	13%					0%
90%	12%	10%	6%				0%	
99%	11%	10%	0%			0%	3%	0%
100%	55%	12%			100%	100%	97%	75%

Table 14: Usable (unsuppressed) cells in case 3

Compared to Cases 1 and 2, more data are made available, as would be expected when the problem is the implied complement.

4.2 Genuine data

In cases 1 and 2, no supression with genuine data led to cells with 1s/2s uncovered. We therefore mostly present cases below with just the ‘uncovered’ bad results.

4.2.1 Case 1

Table 15 shows results for uncovered failures:

Thres-	LFS	Low pay			Charity			
hold	0%	0%	13%	25%	0%	40%	50%	80%
3	x	x			x			
4	x		x		x			
5	x		x		x			
:								
6-15	x			x	x			
:								
20	x			x		x		
25	x		x				x	
30	x	x						x

Table 15: Bad cells (uncovered below threshold) for real data in case 1

In the large dataset (19,000 observations), no cells can be recovered. In the charity datasets (700 observations) very few cells could be recovered until the threshold grows beyond 15; but from 20 onwards the recovery rate is very high for this small dataset and grows with the threshold.

Of most interest is the medium-sized dataset ('low pay': 5,000 observations). The number of problematic cells increases with the threshold, but then appears to stabilise at a threshold of 6. However, the number of bad cells falls as the threshold climbs above 20, which is unexpected given the simulation results.

Table 16 presents the usability data for case 1.

	LFS	Low pay			Charity	
Threshold	100%	81%	94%	100%	90%	95%
3-20	x			x		x
:						
25	x		x			x
30	x	x			x	

Table 16: Usable cells for real data in case 1

Only thresholds above 20 lead to cell suppression. Note that cell suppresion is higher for the medium-sized dataset (low pay) than the small one (charity). This highlights that the charcterstics of the data can be as important for SDC as the bsolute number of observations.

4.2.2 Case 2

For Case 2 results are more complex; see Table 17.

Thres-	LFS		Low pay		Charity					
hold	0%	25%	0%	25%	0%	20%	40%	60%	80%	100%
3	x		x		x					
4	x			x	x					
5	x			x	x					
6	x			x	x					
7	x			x		x				
8	x			x			x			
9	x			x				x		
10	x			x					x	
11	x			x					x	
12	x			x					x	
13	x			x						x
14	x			x						x
15	x			x						x
20	x			x						x
25		x		x						x
30		x	x						x	

Table 17: Proportion of uncovered cells for real data in case 2

For the largest dataset, a higher threshold creates problems where there were none. The smaller LFS dataset does not create a differencing problem at the highest or lowest threshold, but does at all others. For the smallest dataset, there is a positive relationship between the threshold and the number of uncovered rows. The number of usable cells is the mirror image:

	LFS		Low pay			Charity						
Threshold	94%	100%	88%	94%	100%	70%	75%	80%	85%	90%	95%	100%
3		x			x							x
4		x		x								x
5		x		x								x
6		x		x								x
7		x		x							x	
8		x		x						x		
9		x		x					x			
10		x		x				x				
11		x		x				x				
12		x		x				x				
13		x		x			x					
14		x		x			x					
15		x		x			x					
20		x		x			x					
25	x			x			x					
30	x		x			x						

Table 18: Proportion of usable cells for real data in case 2

No cells are suppressed for the large dataset except at the highest thresholds. For the smaller dataset on low pay, 1 cell is suppressed thresholds above 3. The small charity dataset sees cells being suppressed at thresholds above 6, with one-quarter being suppressed at thresholds over 12.

4.2.3 Case 3

This case does produce some uncovered 1s and 2s:

- No failures occur for the large dataset
- 13% of the implicit cells generated by the medium dataset fail
- 60% of the implicit cells generated by the medium dataset fail

For uncovered cells there is more variation, with even the large dataset showing problems:

	LFS			Low pay					Charity		
Threshold	0%	13%	25%	13%	38%	50%	63%	75%	60%	90%	100%
3	x			x					x		
4	x			x						x	
5	x			x						x	
6	x			x							x
7	x			x							x
8	x			x							x
9	x			x							x
10	x			x							x
11	x				x						x
12	x				x						x
13	x					x					x
14	x					x					x
15		x				x					x
20			x			x					x
25			x				x				x
30			x					x		x	

Table 19: Proportion of uncovered cells for real data in case 3

Clearly there are concerns even for the large dataset – but these occur at large thresholds, where the likelihood of a value being below the threshold is higher. For the smaller datasets, the situation is much worse; in the case of the small charity dataset, *all* of the implied complementary dataset is below the threshold when the threshold ranges from 6 to 25. This is highliy likely to be missed by an output checker, particularly as no suppression appears to be need on primary data except for the highest threshold:

	LFS	Low pay		charity	
Threshold	100%	88%	100%	90%	100%
3-25	x		x		x
30	x	x		x	

Table 20: Proportion of usable cells for real data in case 3

4.3 Discussion

The foregoing is an attempt to summary a very large range of statistical findings. The only thing that can be said with certainty are two trivial points that more observations and/or a higher threshold monotonically

- reduces the likelihood of cells with 1 or 2 observations being exposed through differencing or an implied complementarity
- increases the number of suppressed cells

Beyond this, very little can be said definitely, and much of the findings presented above serve to muddy the water.

First, if the purpose of a threshold is to prevent numbers below that threshold being exposed, the monotonic relationships break down. Higher thresholds can lead to more exposure. Even the more-observations-is-good story is no longer clear. In Table 7, for example, a threshold of 30 for 500 observations mens that no errors are found in only 3% of the simulations, but that none of the simulations showed that more than 50% of cells were false suppressions; in contrast, with 5000 observations, 20% of simulations showed no problems but 12% showed that every cell in the implied table was problematic. Which of these is 'better'?

Second, the characteristics of the data are crucial to risk assessment. Table 15 showed that the 5000 observations in the medium dataset generated more problematic cases (in terms of uncovered cells) that the small dataset with under 700 observations. This may be down to the more even split of the 'z' variable in the small dataset; it is hard to tell.

Third, different risk models give different outcomes. Cases 1, 2 and 3 presented quite different results.

Fourth, the risk measure matters. Ensuring no 1s and 2s would seem to be the minimum requirement for an SDC rule. But is 'no uncovered cell' a good rule? If the threshold is 20, then ensuring no implied tble are created with less than 20 units seems to be consistent. But if the purpose of the higher threshold is simply and explicitly to avoid 1s and 2s (as in [12]) then the more general risk measure is not helpful.

Fifth, there is no sense of a natural break in the thresholds. None of the above results suggest that risk probabilities decline notably and consistently after a threshold of 5, or 10, or 20, for example.

Sixth, the proportion of usable cells, as reported above, has very limited value as a measure of retained utility. Guidelines such as [12] emphasise that SDC is applied to small cell values, which should have little or no statistical significance and so their loss is accetble. This may be the case in data such as social surveys, which tend to cluster normally around the centre; but in some fields, such as business or health data,, much of the interest is in the tales and the 'small numbers' being removed by SDC may be crucial for policy inference. A simple count of deleed cells cannot relfect this value.

Finally, the real datasets showed generally fewer problems than the simulated ones. In particular, the uncovering of 1s and 2s (which This is particularly relevant as the small/medium datasets used are comparable in size to the 500/5000 observation simulation reported in section 4.1. Perhaps genuine data is more forgiving than simulations which will, by their nature, generate extreme values. If so, this has implications for using simulations to derive statistical guidelines.

In summary, there is no convincing evidence from this analysis to suggest whether an optimal threshold exists, or even whether this can be measured effectively. We have found no evidence that 10 is a better threshold (in terms of risk management) than any other, or a worse one. In some cases here, 3 performs best and 30 performs worst; in other situations the case is reversed. The only thing that can be said for definite is that retained value (or 'utility') is inversely and monotonically related to the threshold; again, this should not be a surprise.

Hence the choice of a threshold comes down to the institution's comfort level, at the interesection between five related questions:

1. What threshold minimises risk?
2. What threshold generates an acceptable risk?
3. What level of utility loss is acceptable?
4. How will a threshold of x be perceived?
5. Should we have multiple thresholds, and why?

This paper has demonstrated that the answer to the first three is 'we still don't know, and it seems unlikely that we will know'. That leaves questions 4 and 5 as perhaps the main determinant of an institution's threshold. This was certainly the case when the UK Office for National Statistics' secure research facility adopted 10 as the minimum threshold in 2003, one of the earliest to do so. Prior to that point, ONS had required research outputs to have the same threshold of 3 as official statistics, but the research team felt this was too low and didn't demonstrate how output control was being taken seriously. 10 was chosen as one of the researchers was already using it as his personal threshold; a straw poll amongst other lab users suggested this would be acceptable, and a rule was born. However, that rule has migrated into a 'fact': that 10 is a safe number for outputs, whereas it should be abundantly clear from the above that this is not the case.

5 Conclusion

NSIs, and other organisations allowing statistical research on confidential output, need to take a decision on what is an acceptable threshold, irrespective of the output clearance regime.

This paper reports on an attempt to provide some evidence for the particular choice of a threshold. Ultimately this has been unsuccessful; the paper has demonstrated that the relationship between thresholds and risky cells is not linear and depends upon the type of differencing being guarded against, and that differencing measures may have irreconcilable targets.

Some results, not presented here, suggest that as the dataset increases all problems disappear; this is both unsurprising and unhelpful, as the number of observations in a dataset is the maximum of those used in analysis, not a minimum.

On the other hand, when applied to genuine datasets, these results provide some cautious optimism. The largest genuine dataset used in this analysis, with 20,000 observations, is not particularly large by modern NSI standards, and yet it poses almost no differencing risk. Of course, increasing the number of categories would increase the risk potential but, as demonstrated here, the actual impact would depend on the threshold and the measure of 'risk' being used.

One interesting issue is that Case 3 (disclosure by complementarity) seems more problematic than the other cases. This case is discussed in texts such as Hundepool et al (2010), and yet in practice it might be the one most likely to slip under the radar. Moreover, in this case the solution might not be statistical: better guidelines for output checkers and researchers (or enforcement of complete categories by automatic tools) might be more successful than a higher threshold. This re-iterates that SDC is not just a statistical problem[2]; SDC must be part of a coherent system that relfects institutional choices being made.

In terms of further research, there are two areas that might be productive. The simpler is to improve the simulation here, perhaps considering how thresholds might interact with variable distributions, for example. The conceptually easier, but practically much harder, area of research is to exhaustive analyse genuine research publications for actual differencing problems. We are aware of (failed) attempts to do this, and we would welcome collaborations on this important.

Th code and results are available online at http://www.fivesafes.org/SDC/10s_paper/, and the reader is invited to experiment.

References

[1] Ritchie F. and Elliot M. (2015) "Principles- versus rules-based output statistical disclosure control in remote access environments", IASSIST Quarterly v39 pp5-13

[2] Alves, K., & Ritchie, F. (2020). Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control. Statistical Journal of the IAOS, 36(4), 1281-1293. https://doi.org/10.3233/SJI-200661

[3] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. and De Wolf, P-P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC. http://neon.vb.cbs.nl/casc/.\SDC_Handbook.pdf

[4] Serpell M. and Smith J. (2012) Potential Breaches of Confidentiality in Statistical Tables containing Magnitude Data. Proc. 25th European Conference on Operational Research, Data Confidentiality Stream.

[5] Ritchie F. (2014) "Access to sensitive data: satisfying objectives, not constraints", J. Official Statistics v30:3 pp533-545, September. DOI: 10.2478/jos-2014-0033.

[6] Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use, in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki.

[7] Green, E., and Ritchie, F. (2016) Data Access Project: Final Report. Australian Department of Social Services. June

[8] ONS (2007) Default Procedures for Statistical Disclosure Detection and Control v1.1. Mimeo, Office for National Statistics

[9] Skinner C. (2012) Statistical disclosure risk: separating potential and harm. International Statistical Review. 80(3):349–368

[10] Ritchie F. (2019) "Analyzing the disclosure risk of regression coefficients". Transactions on Data Privacy 12:2 (2019) 145 - 173

[11] Green, E., Ritchie, F., Parry, G. and Bradley, P. (2016) Financial resilience in charities v.2. Project Report. University of the West of England. https://uwe-repository.worktribe.com/OutputFile/919303

[12] ONS (2019) Safe Researcher Training [canonical slides, September 2019]. Office for National Statistics.