# 10 is the safest number that there's ever been

Felix Ritchie*

* University of the West of England, Bristol. 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 popular number for both national statistics institute (NSI) outputs and research outputs, five and twenty less so. Some organisations use multiple thresholds for different data sources.

Unfortunately, these are all entirely subjective. Three is the only threshold which has a solid statistical foundation, but many argue that this leaves little margin for error. There is no equivalent statistical case for any larger number: ten is popular because it is popular

This paper tries to provide some empirical analysis by modelling alternative threshold assumptions on both synthetic data and real datasets. The paper demonstrates that there is no 'best' option; moreover, there is no linear relation between a threshold and risk, as higher thresholds can increase disclosure risk in some cases. It also notes that there are disclosure checking practices which can reduce risk irrespective of the threshold.

## 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 Ritchie and Elliott, 2015, 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 fails to achieve either outcome (Alves and Ritchie, 2019). 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 a solid statistical basis, but 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. 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 nice 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, and so ten seemed a very safe suggestion – and moreover, one which was acceptable to researchers. At that time, the decision to use 10 as the threshold by the VML was unusual, and not even common within ONS. Some fifteen years later, ten is the most common number used by it seems appropriate to review this choice again.

This paper tries to provide some empirical analysis of what might be sensible 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.

# 2 Literature review

We are not aware of other literature covering this question.

# 3 Conceptual review

## 3.1 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 use three as the threshold for exposition, as it solves the problem of direct identification.

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 X there exists a second table Y such that (X-Y) has single observations in it. NSIs invest considerable time and effort to ensure that X and Y are not both generated, but this is not a guarantee of protection. Even if Y is not published, how can the NSI guarantee that Y could not be created by some combination of some other tables A, B, C, D…? 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[1]. However, it is 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

[1] NSIs may not review all possible combinations as this is computationally prohibitive in operational circumstances, and this has been shown to be problematic in rare cases.

are also likely to be of little value; it is not at all clear that the same standard needs to be applied to small counts arising from differencing[2].

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' data holders (Ritchie, 2014) 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 (Hafner et al, 2015; Green and Ritchie, 2016) 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.

### 3.2 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 (ONS, 2007) which argued that this gave confidence that simple threshold check would also deal with the problem of multiple respondents from the same

[2] 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. Again, this is a valid argument for a minimum threshold rule, but it does not follow that this should also apply to implicit tables generated through differencing.

business when dealing with hierarchical data. However, a primary motivation for the choice of ten was that it was high enough to avoid questions of differencing but also acceptable to researchers[3].

The VML was not the first such research centre, but since 2003 the number of them has grown steadily, and 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 the UK alone values from five to thirty are used. One organisations uses five as its default, but raises the threshold to ten for more 'sensitive data'. This has the substantial advantage of demonstrating to all concerned 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 (Ritchie, 2019). However, for the threshold rule this is complicated by the apparent absence of any statistical evidence, save for the long-lost analysis of ONS.

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.

# 4 Approach

We tackle this issue by considering three cases which seem to present the most obvious problems. We assume that cell counts of 1 and 2 are values to avoid, irrespective of the formal threshold.

## 4.1 Case 1: differencing between a set and a subset

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

[3] Source: personal discussion.

| 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 higher threshold would have prevented some of the small values, but not all of them. A lower threshold would not have had an effect. Note that we are not concerned that the implicit table has values below the threshold, as this is not what the threshold is designed to achieve. Only the 1s/2s are important.

To consider this:

- Create random category allocation for Age (X)
- Create random urban/rural category allocation for residents (Y) with $p_{urban} > 50\%$
- Create random homeowner/renter category allocation (Z) with $p_{homeowner} > 50\%$
- Tabulate X:Y and X:(Z=homeowner), correcting for the threshold (zero is below threshold)
- Tabulate X:(Z=renter) and count number of 1s/2s in cells where the originals were not suppressed
- Store number of 1s/2s, mean observations and median observations of X:Y and X:(Z=renter)
- Iterate N times with new random values

### 4.2 Case 2: Row totals revealing suppressed cells

Consider Table 3, placed alongside Table 1 for clarity:

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 3 Education

| Age | No degree | Degree | Total |
|---|---|---|---|
| 50-54 | 26 | 6 | 32 |
| 55-59 | 29 | 7 | 36 |
| 60-64 | 36 | <5 | 36 |
| 65+ | 39 | <5 | 39 |
| | 130 | 13 | 143 |

**Tables 1 & 3:** Example of de-suppression through differencing

Although Table 3 has the marginal totals correctly calculated (that is, they add 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, we are looking to uncover suppression rather than find 1s and 2s.

In this case, a lower threshold would have avoided this problem as the 3s and 4s in Table 3 would not have been removed. A slightly higher threshold would not have addressed the problem but a much higher threshold may have as the 'rural' column may 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

### 4.3 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% |

**Tables 1 & 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 a binary conditions being 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 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 (could check against implicit breaches but this would be very onerous and block most outputs unless number of obs is very high; don't test for zero)
- 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 you would not be worried if there are high initial frequencies (w=urban) but might be worried if the initial frequencies awere low (w=rural), so running this way covers both options.

### 4.4 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:

| | 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’:<br>2010 83<br>2011 150<br>2012 151<br>2013 153<br>2014 149 | ‘age’:<br>50-54 6,590<br>55-59 6,366<br>60-64 5,119<br>65-69 957 | ‘age’:<br>50-54 2,091<br>55-59 1,860<br>60-64 850<br>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 (2018). ‘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 (that is, a failure as identified above)

The proportion of ‘ok’ results (that is, the number of usable cells once thresholds had been applied

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.felixritchie.co.uk/sdc_calculations/.

# 5 Results

## 5.1 Simulated data

The simulations produce a very large number of results (2 types of data distributions, 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.

Annex 1 provides samples of results. The log files from running with 500 and 5000 observations are available at the above website, along with a summary spreadsheet for the simulation results.

### 5.1.1 Case 1

As expected, a higher threshold reduces the proportion of 'bad' results (ie where the gap between two non-supressed cells is 1 or 2. As the number of observations increases, the proportion of 'bad' cells falls; see below for thresholds of 3, 10 and 30 (there are 16,000 possible outcomes: 1000 random iterations each assessed at 4Y x 4Z proportions).

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % bad | 3 | 10 | 30 | 3 | 10 | 30 |
| 0% | 815 | 5230 | 8294 | 10480 | 10486 | 12535 |
| 5% | 1537 | 2151 | 1667 | 1548 | 1543 | 1634 |
| 10% | 1726 | 1542 | 1413 | 867 | 870 | 698 |
| 15% | 1787 | 1205 | 1219 | 593 | 589 | 327 |
| 20% | 1950 | 1106 | 1104 | 645 | 645 | 266 |
| 25% | 1826 | 1061 | 1069 | 685 | 687 | 232 |
| 30% | 1524 | 959 | 717 | 583 | 582 | 179 |
| 35% | 1229 | 733 | 392 | 379 | 378 | 90 |
| 40% | 1014 | 571 | 107 | 168 | 168 | 30 |
| 45% | 846 | 486 | 16 | 50 | 50 | 9 |
| 50% | 648 | 352 | 2 | 2 | 2 | 0 |
| 55% | 491 | 276 | 0 | 0 | 0 | 0 |
| 60% | 309 | 161 | 0 | 0 | 0 | 0 |
| 65% | 165 | 103 | 0 | 0 | 0 | 0 |
| 70% | 93 | 45 | 0 | 0 | 0 | 0 |
| 75% | 34 | 18 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 80% | 5 | 1 | 0 | 0 | 0 | 0 |
| 85% | 0 | 0 | 0 | 0 | 0 | 0 |
| 90% | 1 | 0 | 0 | 0 | 0 | 0 |

There is a dramatic difference in the usability of the data. The valid cells left after differencing for Table 2 are:

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % usable | 3 | 10 | 30 | 3 | 10 | 30 |
| 0% | 0 | 0 | 160 | 0 | 0 | 0 |
| 5% | 0 | 0 | 401 | 0 | 0 | 0 |
| 10% | 0 | 0 | 577 | 0 | 0 | 0 |
| 15% | 0 | 0 | 707 | 0 | 0 | 0 |
| 20% | 0 | 0 | 783 | 0 | 0 | 0 |
| 25% | 0 | 0 | 953 | 0 | 0 | 0 |
| 30% | 0 | 0 | 1378 | 0 | 0 | 0 |
| 35% | 0 | 0 | 1801 | 0 | 0 | 0 |
| 40% | 0 | 0 | 1989 | 0 | 0 | 0 |
| 45% | 0 | 0 | 2499 | 0 | 0 | 0 |
| 50% | 100 | 7689 | 4752 | 0 | 0 | 2508 |
| 55% | 344 | 994 | 0 | 0 | 0 | 959 |
| 60% | 711 | 739 | 0 | 0 | 0 | 416 |
| 65% | 955 | 802 | 0 | 0 | 0 | 93 |
| 70% | 908 | 828 | 0 | 0 | 0 | 23 |
| 75% | 809 | 803 | 0 | 0 | 0 | 23 |
| 80% | 742 | 793 | 0 | 0 | 0 | 56 |
| 85% | 961 | 772 | 0 | 0 | 2 | 164 |
| 90% | 1133 | 887 | 0 | 0 | 19 | 387 |
| 95% | 1383 | 1002 | 0 | 0 | 235 | 679 |
| 100% | 7954 | 691 | 0 | 16000 | 15744 | 10692 |

### 5.1.2 Case 2

For this case, there are 12,000 outcomes (1000 iterations by 4Y and 3 Q proportions).

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 de recovered) 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 threshold of 30 is considerably worse. The result seems to be because, as the number of observations increases, a higher threshold increase sthe chance that one or other row (but not both) has just one cell suppressed.

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % bad | 3 | 10 | 30 | 3 | 10 | 30 |
| 0% | 2475 | 1363 | 5011 | 12000 | 11997 | 6005 |
| 10% | 1908 | 1806 | 3782 | 0 | 3 | 149 |
| 20% | 1466 | 1520 | 2101 | 0 | 0 | 242 |
| 30% | 1366 | 1143 | 833 | 0 | 0 | 295 |
| 40% | 1551 | 936 | 228 | 0 | 0 | 195 |
| 50% | 1484 | 973 | 43 | 0 | 0 | 178 |
| 60% | 971 | 990 | 2 | 0 | 0 | 398 |
| 70% | 559 | 817 | 0 | 0 | 0 | 894 |
| 80% | 191 | 862 | 0 | 0 | 0 | 1503 |
| 90% | 27 | 939 | 0 | 0 | 0 | 1486 |
| 100% | 2 | 651 | 0 | 0 | 0 | 655 |

The results on usable cells are also complex. With a threshold of 10, in none of the 1000 iterations do more 75% of the rows retain valid values; with 5,000 observations, only 28 iterations did not leave all values unsuppressed:

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % usable | 3 | 10 | 30 | 3 | 10 | 30 |
| 40% | 0 | 0 | 60 | 0 | 0 | 0 |
| 45% | 0 | 0 | 1076 | 0 | 0 | 0 |
| 50% | 88 | 5936 | 8396 | 0 | 0 | 952 |
| 55% | 384 | 3560 | 1776 | 0 | 0 | 1320 |
| 60% | 840 | 1740 | 512 | 0 | 0 | 1144 |
| 65% | 1104 | 612 | 140 | 0 | 0 | 444 |
| 70% | 912 | 128 | 36 | 0 | 0 | 128 |
| 75% | 692 | 24 | 4 | 0 | 0 | 12 |
| 80% | 840 | 0 | 0 | 0 | 0 | 0 |
| 85% | 1156 | 0 | 0 | 0 | 0 | 0 |
| 90% | 1560 | 0 | 0 | 0 | 0 | 4 |
| 95% | 2104 | 0 | 0 | 0 | 28 | 72 |
| 100% | 2320 | 0 | 0 | 12000 | 11972 | 7924 |

### 5.1.3 Case 3

As for Case 2, there are 12,000 outcomes (1000 iterations, 4Y and 3W proportions).

When considering the risk in binary complements, there appears to be a large risk even with a threshold of 10 when the number of observations is small. More interestingly, increasing the number of observations has a much larger impact brings results for the lower thresholds very much in line with the higher ones.

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % bad | 3 | 10 | 30 | 3 | 10 | 30 |
| 0% | 81 | 409 | 687 | 3392 | 3392 | 4382 |
| 5% | 328 | 894 | 1596 | 1071 | 1071 | 1999 |
| 10% | 792 | 1252 | 2135 | 835 | 835 | 1159 |
| 15% | 1238 | 1507 | 2219 | 882 | 882 | 692 |
| 20% | 1613 | 1773 | 2091 | 1087 | 1087 | 581 |
| 25% | 1548 | 1539 | 1568 | 1177 | 1177 | 591 |
| 30% | 1383 | 1265 | 1038 | 1058 | 1059 | 591 |
| 35% | 1185 | 898 | 485 | 845 | 845 | 561 |
| 40% | 1061 | 682 | 152 | 614 | 613 | 462 |
| 45% | 896 | 563 | 27 | 449 | 449 | 396 |
| 50% | 704 | 455 | 2 | 278 | 278 | 274 |
| 55% | 532 | 351 | 0 | 196 | 196 | 196 |
| 60% | 331 | 219 | 0 | 75 | 75 | 75 |
| 65% | 189 | 120 | 0 | 26 | 26 | 26 |
| 70% | 74 | 40 | 0 | 13 | 13 | 13 |
| 75% | 35 | 27 | 0 | 2 | 2 | 2 |
| 80% | 8 | 4 | 0 | 0 | 0 | 0 |
| 85% | 2 | 2 | 0 | 0 | 0 | 0 |

More observations does increase the number of usable cells, but there remains a large information loss associated with the higher threshold :

| | 500 observations | | | 5000 observations | | |
|---|---|---|---|---|---|---|
| % usable | 3 | 10 | 30 | 3 | 10 | 30 |
| 25% | 0 | 0 | 3 | 0 | 0 | 0 |
| 30% | 0 | 0 | 120 | 0 | 0 | 0 |
| 35% | 0 | 0 | 465 | 0 | 0 | 0 |
| 40% | 0 | 0 | 1179 | 0 | 0 | 0 |
| 45% | 0 | 0 | 2007 | 0 | 0 | 0 |
| 50% | 9 | 5169 | 8226 | 0 | 0 | 441 |
| 55% | 72 | 714 | 0 | 0 | 0 | 831 |
| 60% | 213 | 171 | 0 | 0 | 0 | 930 |
| 65% | 519 | 210 | 0 | 0 | 0 | 543 |
| 70% | 666 | 510 | 0 | 0 | 0 | 177 |
| 75% | 675 | 732 | 0 | 0 | 0 | 75 |
| 80% | 582 | 795 | 0 | 0 | 0 | 3 |
| 85% | 540 | 525 | 0 | 0 | 0 | 0 |
| 90% | 837 | 519 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 95% | 1200 | 1113 | 0 | 0 | 3 | 27 |
| 100% | 6687 | 1542 | 0 | 12000 | 11997 | 8973 |

## 5.2 Genuine data

In this section, we present simply the "none", "some", "all" findings.

### 5.2.1 Case 1

With the genuine data, Case 1 presents no problems for any dataset. There were no bad cells that could have been recovered. There were some suppressed cells for the smaller datasets (low pay at high thresholds, charity data across the board).

| | Usable cells | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LFS | | | | Low pay | | | | Charity | | | |
| | 75% | 88% | 90% | 100% | 75% | 88% | 90% | 100% | 75% | 88% | 90% | 100% |
| Threshold | | | | | | | | | | | | |
| 3 | | | | 1 | | | | 1 | | | 1 | |
| 4 | | | | 1 | | | | 1 | | | 1 | |
| 5 | | | | 1 | | | | 1 | | | 1 | |
| 6 | | | | 1 | | | | 1 | | | 1 | |
| 7 | | | | 1 | | | | 1 | | | 1 | |
| 8 | | | | 1 | | | | 1 | | | 1 | |
| 9 | | | | 1 | | | | 1 | | | 1 | |
| 10 | | | | 1 | | | | 1 | | | 1 | |
| 11 | | | | 1 | | | | 1 | | | 1 | |
| 12 | | | | 1 | | | | 1 | | | 1 | |
| 13 | | | | 1 | | | | 1 | | | 1 | |
| 14 | | | | 1 | | | | 1 | | | 1 | |
| 15 | | | | 1 | | | | 1 | | | 1 | |
| 20 | | | | 1 | | | | 1 | | | 1 | |
| 25 | | | | 1 | | | 1 | | | | 1 | |
| 30 | | | | 1 | 1 | | | | | | 1 | |

### 5.2.2 Case 2

For Case 2 results are more mixed. 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 at-risk rows.

| | LFS | | | Low pay | | | Charity | | |
|---|---|---|---|---|---|---|---|---|---|
| Risky cells: | none | some | all | none | some | all | none | some | all |

| Threshold | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | | | 1 | | | 1 | | | |
| 4 | 1 | | | | 1 | | 1 | | | |
| 5 | 1 | | | | 1 | | 1 | | | |
| 6 | 1 | | | | 1 | | 1 | | | |
| 7 | 1 | | | | 1 | | | 1 | | |
| 8 | 1 | | | | 1 | | | 1 | | |
| 9 | 1 | | | | 1 | | | 1 | | |
| 10 | 1 | | | | 1 | | | 1 | | |
| 11 | 1 | | | | 1 | | | 1 | | |
| 12 | 1 | | | | 1 | | | 1 | | |
| 13 | 1 | | | | 1 | | | | 1 | |
| 14 | 1 | | | | 1 | | | | 1 | |
| 15 | 1 | | | | 1 | | | | 1 | |
| 20 | 1 | | | | 1 | | | | 1 | |
| 25 | | 1 | | | 1 | | | | 1 | |
| 30 | | 1 | | 1 | | | | 1 | | |

No cells are suppressed for the LFS data except at the highest thresholds. For the smaller LFS dataset on low pay, 1 cell is suppressed at all thresholds. The small charity dataset sees cells being suppressed at thresholds above 5, with half the cells being suppressed at a threshold over 11.

| | LFS | | Low pay | Charity | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 88% | 100% | 88% | 50% | 60% | 70% | 80% | 88% | 90% | 100% |
| Threshold | | | | | | | | | | |
| 3 | | 1 | 1 | | | | | | | 1 |
| 4 | | 1 | 1 | | | | | | | 1 |
| 5 | | 1 | 1 | | | | | | | 1 |
| 6 | | 1 | 1 | | | | | | 1 | |
| 7 | | 1 | 1 | | | | 1 | | | |
| 8 | | 1 | 1 | | | 1 | | | | |
| 9 | | 1 | 1 | | 1 | | | | | |
| 10 | | 1 | 1 | | 1 | | | | | |
| 11 | | 1 | 1 | | 1 | | | | | |
| 12 | | 1 | 1 | 1 | | | | | | |
| 13 | | 1 | 1 | 1 | | | | | | |
| 14 | | 1 | 1 | 1 | | | | | | |
| 15 | | 1 | 1 | 1 | | | | | | |
| 20 | 1 | | 1 | 1 | | | | | | |
| 25 | 1 | | 1 | 1 | | | | | | |

| | | | |
|---|---|---|---|
| 30 | 1 | 1 | 1 |

### 5.2.3 Case 3

For the full LFS dataset, there are no suppressed cells and no differencing problems. For the other two datasets, there are some complementarity problems at every threshold, despite cells being suppressed only at a threshold of 30.

### 5.3 Discussion

The foregoing is an attempt to summary a very large range of statistical findings. Some general points can be brought out.

First, as a general rule a higher threshold does provide a higher level of protection. However, it can also remove a substantial amount of useful information, even with large datasets.

Second, this conclusion varies with the type of problem being solved. Simulation Case 2 shows that the relationship between threshold and risk is concave; moreover, adding more observations improves the performance of lower thresholds for both risk and value, whereas for high thresholds it worsens risk without the expected gain in performance.

Third, adding more observations does not necessarily improve outcomes. The negative performance for high thresholds in Case 2 persists with 10,000 observations (higher numbers not tested yet).

Finally, for genuine data the usual differencing problem described as Case does not present realistic problems, even in small datasets. Case 2 does create differencing opportunities in the smaller datasets, but again the relationship with the threshold is non-linear. Case 3 present problems for the smaller datasets; this may be missed by an output checker as few cells in the ‘main’ table are suppressed, indicating plenty of observations.

## 6 Conclusion

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 in practice the number of observations in a dataset is a maximum, not a minimum.

On the other hand, when applied to genuine datasets, these results provide some cautious optimism. The largest real dataset, with 20,000 observations, is not particularly large by modern 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.

This paper provides little 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 value is inversely and monotonically related to the threshold; again, this should not be a surprise.

One interesting issue is that Case 3 (disclosure by complementarity) seems more problematic than the other cases. This case seems to be rarely discussed in texts, and yet it might be the one most likely to slip under the radar. This might be an area worth exploring further, although the solution might be better guidelines for output checkers and researchers rather than a higher threshold.

All of the code and results are available online, and the reader is invited to experiment[4].

## References


Alves K. and Ritchie F. (2019) *Runners, repeaters, strangers and aliens: operationalising efficient output disclosure control*. Working papers in Economics no 1904, University of the West of England.

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

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.

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

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.

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


[4] And to let the author know of coding errors, in a pleasantly non-judgmental manner.

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

## Annex 1 Sample outputs

Each analysis produces three types of outputs:

**The proportion of bad cells.**

The example below is taken from Case 1, 500 observations, 1000 iterations, Y=70% Z=70%. The table shows that in 145 simulations out of 1000, none of the cells generated a re-identification problem when the threshold was 3; 997 out of 1000 had no problems when the threshold was 30. When the threshold was 3-8, 1 simulation out of 1000 always showed that 30% of the suppressed cells were problematic.

| Cells | Threshold | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| % bad | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 20 | 25 | 30 |
| 0% | 145 | 147 | 150 | 156 | 172 | 209 | 252 | 321 | 421 | 525 | 638 | 726 | 810 | 973 | 991 | 997 |
| 5% | 369 | 370 | 374 | 379 | 389 | 396 | 422 | 427 | 386 | 344 | 288 | 238 | 174 | 27 | 9 | 3 |
| 10% | 274 | 271 | 269 | 267 | 261 | 245 | 222 | 177 | 148 | 103 | 62 | 29 | 14 | 0 | 0 | 0 |
| 15% | 153 | 155 | 152 | 149 | 131 | 116 | 83 | 60 | 38 | 25 | 11 | 6 | 2 | 0 | 0 | 0 |
| 20% | 42 | 40 | 40 | 35 | 34 | 25 | 14 | 13 | 7 | 3 | 1 | 1 | 0 | 0 | 0 | 0 |
| 25% | 16 | 16 | 14 | 13 | 12 | 8 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30% | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**The proportion of usable cells in each source table**

This table shows how many cells out of the totals were not suppressed (a zero value was counted as not-surpressed). For the same data iteration, for Table 1 and 2. The table shows for, example, that

- With a threshold of 3, no cells were suppressed in the original table in any iteration; 1 cell was suppressed 8 times (20 possible table cells, with 95% not suppressed)
- With a threshold of 30, no more than 50% of cells in Table 1 were unsuppressed, and in Table 2at most 25% of cells were let unsuppressed

Table 1

| | Threshold | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ok% | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **20** | **25** | **30** |
| 25% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 30% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 40 |
| 35% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 153 |
| 40% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 350 |
| 45% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 199 | 340 |
| 50% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 269 | 707 | 116 |
| 55% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 396 | 66 | 0 |
| 60% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 251 | 1 | 0 |
| 65% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 11 | 71 | 69 | 0 | 0 |
| 70% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 21 | 76 | 189 | 13 | 0 | 0 |
| 75% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 52 | 164 | 250 | 1 | 0 | 0 |
| 80% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 13 | 69 | 163 | 247 | 234 | 0 | 0 | 0 |
| 85% | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 16 | 75 | 176 | 290 | 267 | 163 | 0 | 0 | 0 |
| 90% | 0 | 0 | 0 | 0 | 3 | 7 | 31 | 108 | 209 | 290 | 264 | 169 | 66 | 0 | 0 | 0 |
| 95% | 0 | 1 | 7 | 24 | 53 | 127 | 234 | 360 | 401 | 320 | 167 | 58 | 7 | 0 | 0 | 0 |
| 100% | 1,000 | 999 | 993 | 976 | 944 | 866 | 732 | 513 | 301 | 132 | 40 | 6 | 0 | 0 | 0 | 0 |

Table 2

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 156 |
| 5% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 345 |
| 10% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 327 |
| 15% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 115 | 145 |
| 20% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 225 | 25 |
| 25% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 305 | 2 |
| 30% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 216 | 0 |
| 35% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 121 | 88 | 0 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 327 | 17 | 0 |
| 45% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 13 | 36 | 355 | 3 | 0 |
| 50% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 8 | 52 | 158 | 332 | 163 | 1 | 0 |
| 55% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 13 | 54 | 178 | 321 | 387 | 8 | 0 | 0 |
| 60% | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 51 | 166 | 315 | 318 | 188 | 0 | 0 | 0 |
| 65% | 0 | 0 | 0 | 0 | 0 | 1 | 7 | 30 | 126 | 250 | 244 | 129 | 43 | 0 | 0 | 0 |
| 70% | 0 | 0 | 0 | 0 | 1 | 1 | 13 | 95 | 244 | 275 | 148 | 47 | 10 | 0 | 0 | 0 |
| 75% | 0 | 0 | 0 | 0 | 1 | 10 | 74 | 209 | 255 | 162 | 53 | 13 | 2 | 0 | 0 | 0 |
| 80% | 0 | 0 | 0 | 0 | 7 | 48 | 180 | 278 | 189 | 68 | 8 | 1 | 0 | 0 | 0 | 0 |
| 85% | 0 | 0 | 0 | 4 | 42 | 192 | 277 | 227 | 84 | 14 | 1 | 0 | 0 | 0 | 0 | 0 |
| 90% | 0 | 0 | 9 | 57 | 206 | 311 | 279 | 115 | 30 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 95% | 8 | 52 | 151 | 318 | 380 | 303 | 134 | 31 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 100% | 992 | 948 | 840 | 621 | 363 | 134 | 36 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Overall perspective on problems**

The third table automatically produced summaries the proportion of problematic cells into a simple sum yes/no count; see below. The table below shows that, with a threshold of 3, every iteration produced at least one problematic cell. One in a thousand iterations turned a case with no problems when the threshold was 4 or 5. But even when the threshold was 30 in 92 cases at least one cell would create a differencing problem between Table 1 and Table 2.

| Threshold | none | some |
|---|---|---|
| 3 | 0 | 1,000 |
| 4 | 1 | 999 |
| 5 | 1 | 999 |
| 6 | 3 | 997 |
| 7 | 13 | 987 |
| 8 | 25 | 975 |
| 9 | 73 | 927 |
| 10 | 169 | 831 |
| 11 | 301 | 699 |
| 12 | 469 | 531 |
| 13 | 620 | 380 |
| 14 | 719 | 281 |
| 15 | 791 | 209 |
| 20 | 872 | 128 |
| 25 | 881 | 119 |
| 30 | 908 | 92 |