The SACRO guide to statistical output checking

SACRO Workpackage 1 Deliverable

October 2023

About this document

This document was developed as an output of the project SACRO (Semi-Automatic checking of Research Outputs). This work is funded by UK research and Innovation [Grant Number MC_PC_23006] as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, delivered in partnership with Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK).

The conceptual development was largely carried in workshops and discussions within the SACRO project team (Elizabeth Green, Felix Ritchie, Jim Smith, Amy Tilbrook and Paul White). The primary author is Felix Ritchie (contact: felix.ritchie@uwe.ac.uk), with additional text and extensive comments supplied by other members of the SACRO project team, particularly on the statistical elements. Cara Kendal produced the summaries of control measures, other guides and references. Other members of the SACRO network and external organisations provided additional feedback, particularly on the way that the model was presented to non-technical audiences.

The statbarn model was first presented on 18th May to the SACRO network, and subsequently to the UNECE/Eurostat Expert Group on Statistical Data Confidentiality, Wiesbaden, September 2023. This completed document will be circulated amongst the community for consultation, used to review UK training practices, and form the basis of a workshop in Spring 2024. The final version will be published in March 2024.

For information on developments, see the project website https://medium.com/sacro-semi-automated-checking-of-research-outputs. Comments on this document are welcomed – for an editable version of this document, or to discuss the guide or its concepts, please contact the project team.

Citation: SACRO (2023) SACRO Guide to Statistical Output Checking. Project report. Bristol, October.

10.5281/zenodo.10054629

October 2023

Table of contents

Part I Operational and Statistical theory

1. Introduction

1.1 Document structure

This document has three sections, plus references:

- **Part I** discusses the operational and statistical theory underlying all aspects of output statistical disclosure control (OSDC). It is intended for those who want a deeper understanding of the concepts underlying practical and efficient OSDC.
- **Part II** is the practical manual, intended to be a reference for users and output checkers. It lists the rules (or rules-of-thumb) to be followed for all possible outputs, organised by class. For each class of statistics, the manual presents a quick summary, an example, a discussion of the key risks, a rationale for any rules applied, and guidelines on how to evaluate the output if an exception is requested. A detailed statistical explanation of the risk assessment is not given, but other works are referenced for those interested. A separate chapter discusses graphs.
- **Part III** provides FAQs on both statistics and outputs, plus some guidance for output checkers on how to respond to queries

The expectation for the guide is that Parts II and III will be regularly updated as our understanding of SDC develops and as new FAQs arise; Part I should be fairly stable.

1.2 Audience

This manual is intended to comprehensively support researchers, output checkers, and data governance managers.

- **Researchers** need only consider Part II, perhaps part I on safe statistics, perhaps the FAQs; however, they may find the rest of the manual enlightening
- **Output checkers** should read the whole manual, and have Part II to hand when checking. **Senior output checkers** should be familiar with the whole manual, and follow up the references; in particular Ritchie and Welpton (2015) and Alves and Ritchie (2021).
- **Data governance managers** may skip Part II but should study Parts I and III, and follow up the references for senior output checkers

1.3 Output checking in context

Output checking is one part of secure and efficient management for confidential data. The most widely-used framework for confidential research data governance is the Five Safes (Ritchie, 2017). This proposes five dimensions of ‘control’ or ‘risk’ that needs to be considered when planning research data management:

- Projects – why is this project being done, who for, what happens at the end?
- People – will those with access to data look after it, and do they need training?
- Setting – will the physical/technical environment limit the chance of accidental or deliberate misuse?
- Data – is the level of data appropriate for the project?
- Outputs – is there any residual risk in publication?

Whilst the first four are often thought about, outputs are usually the forgotten relations. Training (or even awareness that this is a problem) is currently limited to those using secure research environments, although this is likely to change in the future.

The reason that it can be forgotten is because statistical research outputs are generally very low risk – good statistics are generally (although not always) aligned with negligible disclosure. The converse is true: disclosure often is associated with poor statistics, and so SDC can be seen as supporting the researcher.

SDC is therefore not the only protection applied to confidential data. There are usually other checks to make sure the researchers are competent and aware of their duty of confidentiality. Note that we assume any SDC problem arise from poor statistics or mistakes by the researcher; unlike the other four safes, output checking does not consider 'bad hat' actors. We do not allow for a researcher deliberately falsifying results to breach confidentiality because (a) it probably can't be detected and (b) the 'people' dimension should be ensuring that only people who will work genuinely have access to the data.

1.4 Basic OSDC terms

These terms will be used throughout the guide.

Threshold	The minimum number of observations that a statistic should be based on. Can apply to all statistics, although generally not used for models. The purpose is to prevent disclosure from a statistic based on one or two observations. Typically set by the data owner. 3 is used in textbooks as it is the only value with a clear statistical justification. Most organisations now use a higher value (5 or 10) to (a) provide a margin of error and (b) to protect against differencing. This is an organisational preference. See Ritchie (2021) for a discussion
(Residual) degrees of freedom	Generally used as the equivalent threshold for models and test statistics. The purpose is to prevent an equation from masquerading as a model. Residual degrees of freedom is, broadly speaking, number of observations minus number of restrictions embodied in the test or modelled eg for a simple linear regression it is N-K, where K is the number of coefficients including the intercept; for a chi-square it would be N-the test degrees of freedom. Since Brandt et al (2010), 10 is widely used as the minimum.
Dominance	Relevant for magnitudes, but not for frequencies. Dominance checks are there to prevent disclosure where a statistic, although seemingly with many observations, in practice is largely determined by one or two values (for example, mean earnings in a small village where one resident is a Premiership footballer). A dominant observation means may be guessable to a reasonable degree of approximation. Dominance tests vary with the statistic. Dominance is hard to spot as there is no clue from the statistic; this makes it harder to check, but also harder for an attacker, who is required to have a very high level of knowledge. Note that Brandt et al (2010) and derived texts discuss 'row dominance', where much of the value of a table row or column is concentrated in a single cell. This is a statistical problem, but not a disclosure issue.
Class disclosure	This occurs when everyone in a defined set has a defined characteristic eg "all of the students surveyed had taken cannabis"; "no-one in the village earns over £45,000 per year". If the characteristic (drug use, earnings) is informative about members, **and** if membership of the set can be reasonably determined (all students in the survey; everyone in the village) then this is a disclosure; but both of these are highly context-sensitive.
Structural or evidential zeros	A class disclosure does not occur where we would expect that class characteristic to be true or false eg "none of the patients aged 80+ had a living parent". Note: the literature usually only refers to structural zeros; see Appendix 3 for a more detailed discussion of why the distinction matters

Differencing	Differencing refers to the disclosure risk that may occur when two statistics are based on the same data but one has one or two additional observations. Potentially, the difference between the two statistics creates an implicit statistic about the additional observation. For example, mean earnings in a village of 16 people are £49,125; the researcher decides to omit the highest earner as an outlier and calculates a mean of £32,400 for the remaining 15. The difference shows that the person omitted earned £300,000.
Model saturation	Module saturation occurs when all possible combinations of variables are included as explanatory factors. This can lead to a table masquerading as a model. Only likely to be an issue with one or two binary variables eg with three binary variables there are eight possible combinations that must all be included in the model. Note that this is unrelated to saturation when it is used to indicate that adding more data makes no difference to results.

Table 1 Terms used in the guide

2. Risk assessment

The risks in any output need to be assessed before the output is released. How this is done can have a large effect on the clearance process. What makes an output ‘disclosive’? There is no agreed measure. Statistical tools can generate models of re-identification risk, but generally these are meaningless as they are unable to model the outside information of a potential attacker (Hafner et al., 2016).

There are two options:

- To refuse to release outputs until there are demonstrably no meaningful risks
- To release outputs unless there are demonstrably meaningful risks

Although in theory these produce the same outcomes, in practice these two approaches (default-closed and default-open, respectively) differ. Even if they give the same outcome, the default-closed approach is likely to be more time consuming, and to require more engagement with researchers.

The negative statement “we can’t prove that the data could not be re-identified” is almost certainly true; equally, it is meaningless because there is an infinity of possibilities to be disproven. However many times a person shows why X does not lead to re-identification, then the argument changes to “well, how do we know that Y can’t lead to re-identification?” This is the ‘what if..?’ approach, and is often used by those unwilling to commit and hoping to have decisions taken out of their hands. Using the ‘what-if..?’ approach, it is easy to devise potential scenarios even if not plausible, and demand that that are all refuted before action can be taken.

A better (and more practical) way to address this question is to

(1) assume there is no re-identification risk
(2) identify specific circumstances which could lead to a breach, and assess their plausibility

We refer to this as ‘chain-of-events’ reasoning – what chain of events would plausibly lead to the outcome in question occurring, and is that chain of events likely or relevant?

Consider the example in Figure 1, taken from the DRAGoN Output Checking Course.

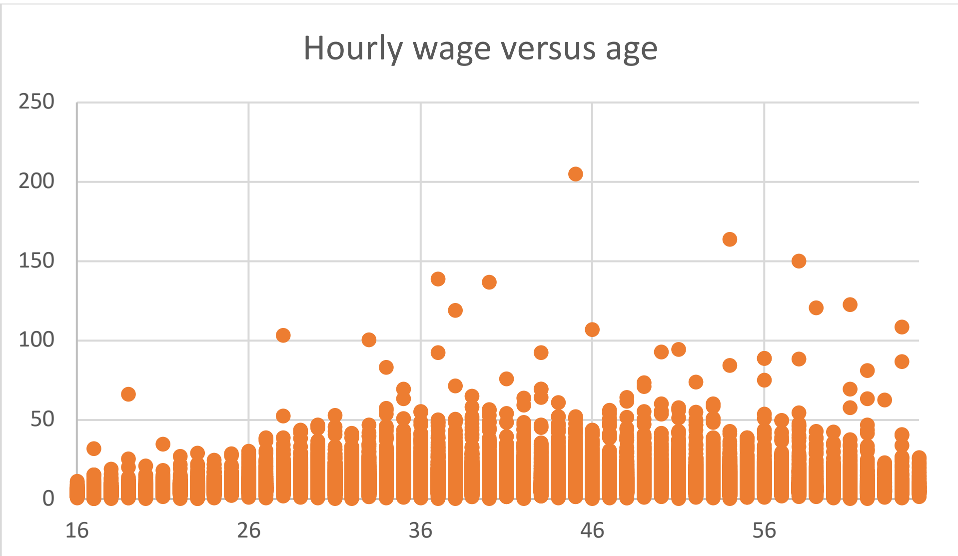


Figure 1 Example scatter plot, hourly wage versus age

We consider two risks: that the outlier for the 45-year-old (or any other group) is identifiable, and that the maximum wages at any age group could be informative.

For the 45 year-old outlier, there are a large number of observations with the same characteristic. What specific scenarios would lead to the person being identified? Some ideas:

a) the data comes from a known population and that the highest earner is known but the exact value is not known (eg all data from one company, the highest earner being CEO)
b) other information in the report provides missing demographic data ("all the highest earners are lawyers", "all the highest earners live in West London")
c) the highest earner told a friend that she took part in the survey, and suggests that the friend check it out to see if he can find her
d) the individual self-identifies by seeking out the survey results

Scenario (a) is possible but easily checked. Scenarios (b) and (c) are possible but seem highly unlikely in reality. Scenario (d) is more feasible, but given that no new information is created (the respondent knows already what she earns, which is why she is able to self-identify), is this a disclosure? Self-identification can be problematic, but in general (and in much of data protection law) this is not a significant problem as it extremely difficult to show that it is not possible. In summary, the outlier does not appear to be breaching confidentiality.

We could consider a case for the outlier for 19 year-olds – one of whom appears to be earning about £70 an hour, a very high value especially as this data is from 2002. Perhaps this is an actor, a footballer, an internet pioneer? The default-closed process requires us to exclude these possibilities. Taking a default-open approach, we acknowledge the existence of such scenarios as feasible – but recognise that there is very little likelihood of that individual being uncovered.

Finally, we consider the maxima displayed. Technically, these are class disclosures as we know the maximum earnings for each age. But are these practical disclosures? The answer hinges on whether it seems likely that "no-one aged X earns more than Y" is informative. Even the 16 year olds have at least one person earning £20/hour, which translates into £36,000 pa – twice the mean adult wage in 2002. For most other ages, the maximum observed wage is two or three times this. So the maxima tell us that "no-one in this sample earns millionaire-style wages, but people in each age group can earn a very good wage". This does not seem informative.

Hence, without further information coming to light, there seems no disclosure risk with this graph, and it can be released safely.

However, even where there is no *disclosure* risk, there may be a *reputational* risk. A naïve reader of the graph could consider that the outliers are evidently disclosive and complain, in which case the data owner is on the back foot to respond. Note that the accuracy of the identification is not a concern. It is the belief that this is disclosive that causes the problems. Hence, risk assessment can consider whether an output is likely to give a misleading impression of disclosure.

Consider this table, reporting on an (imaginary) vox pop taken on the streets of Cardiff

	Views on independence		
	Pro	Anti	None
Welsh speaker	14	4	9
Non-Welsh speaker	1	12	6

Table 2 Small numbers example for reputation risk

Although there is little to identify participants, the single observation (Not Welsh speaker, pro-independence) in a known area could be picked up as “That must be X that you talked to, then”. This is more likely to happen in tables, where the single values are obvious. Hence, this is one reason why small numbers are avoided even if non-disclosive.

The discussion above focuses on perfect data. However, it should be clear that other factors need to be considered. Factors increasing disclosure risk include known population, sample, geographical area, or timing of the data collection. Factors reducing disclosure risk include low data precision, low data quality and missing values, particularly if this is noted in the output (“only 65 of the sample of 75 provided usable data”).

The default-open model is generally seen as good practice (Green, Ritchie, Tava et al., 2021); as well as being more consistent with the aim of output checking (get as much output released subject to addressing disclosure risks), it reflects that research outputs are very low risk to start with.

3. Classification of statistics (a): ‘safe’ and ‘unsafe’

The concept of ‘safe statistics’ (Ritchie, 2007) a precursor to the statbarn taxonomy to be introduced in the next section. Some statistics, such as tables of frequencies, present a number of potential disclosure risks. Others, such as the coefficients of correlation from model estimates, have no effective risk. The concept of ‘safe’ versus ‘unsafe’ statistics helps to clarify this.

An ‘unsafe’ statistic has meaningful disclosure risks, and so those risks need to be assessed before the output can be approved for public circulation. The specific instance of the statistic is assessed.

A ‘safe’ statistic has no meaningful disclosure risk, and so can be published without further checks except administrative ones. The definition of a safe statistic is based on its mathematical form, not its statistical meaning, and so this is independent of whether the data used to generate it are sensitive or not; for a safe statistic, it is not important to know anything about the underlying data. Even if the administrative checks are not done, the statistic is very unlikely to produce a disclosive output and can be largely ignored in the output checking process.

Note that in training (eg Safe Researcher Training), safe and unsafe statistics are usually referred to as ‘low review’ and ‘high review’ statistics, respectively. They may also be presented as if on a spectrum, rather than a binary classification, to show differing levels of risk in ‘unsafe statistics.

3.1 Principles of ‘safe/unsafe’ classification

The key principles can be stated

- Statistics are generally assumed to be unsafe
- A statistic can be classified as safe if there are demonstrably no disclosure risks, irrespective of the characteristics of the data
- A safe statistic may require some checks as best practice, but omitting these does cause unreasonable risks in any genuine research environment
- Unlikely combinations of variables generated specifically to show disclosiveness of safe statistics are not valid counter-arguments to the classification

Note that these principles do *not* say that no checks are required; only that these are best practice. But the classification does not depend on these being carried out, as we would expect that there is no meaningful disclosure risk in genuine use.

For example, the best-practice rules for regressions require checking (a) that the number of residual degrees of freedom exceed some threshold, and (b) that the regression is not fully saturated (ie all possible combinations of regressors included as explanatory variables). A genuine researcher would need to be both extraordinarily incompetent and energetic to achieve this. Even in the only plausible case, of simple linear regression on a single binary variable or two interacted binary variables, the researcher generates a table of means which *could* be disclosive is one category has only a single observation. Using a default-open approach to risk assessment, this is extended chain of events that lead to a non-trivial disclosure can be considered negligible.

As a second example, consider the Herfindahl index. The best-practice rule requires checking that the root of the index is some distance away for the share of the largest value; as a safe statistic tests, this check does not require any knowledge or understanding of the data but is a mathematical test. As with all dominance measures, this can only be checked at the point of calculation, not from the statistic itself. What happens if the check is carried out at the root index does indeed approximate the share of the largest value (implying all the other observations have negligible share)? An intruder may suspect that the root index reflects the largest share, but with a degree of uncertainty. At best, the intruder can state with certainty the upper bound of the largest contributor's share (ie root index). For this to be useful, the total value from which shares are calculated needs to be available, but this is not necessarily the case. We also need to think how the Herfindahl index is used. A common use is to estimate monopoly power in business statistics, where the expectation is that there are competitors; otherwise the estimation is trivial. Another example of use is to look at how the respondent distributes some activity or resource between alternatives (for example, as in Green, Ritchie Bradley and Parry, 2021). In this case, an individual value is of no interest, even at its minimum or maximum. Hence we assess the index as 'safe'.

As third example, weighted frequencies might appear to have very little risk, particularly if the weights are unknown. However, we assume that the weights may be available (this is very reasonable if the data uses standard weights supplied by the data owner, such as statistical offices do); moreover, we can envisage situations where the weights are the same across all cases, meaning the weighted statistic is a linear transformation of the unweighted one. While these may be unlikely cases, the fact that they exist and are not negligible implies they need to be checked; hence, weighted frequencies are classed as 'unsafe'.

3.2 Determining a safe/unsafe classification

Ritchie (2014) proposed a process for considering whether a statistic should be classed as safe:

1. Describe the 'base' (highest risk) form of the statistic – usually the simplest version; for example, a linear regression with no iterative components and no incidental parameters.
2. Determine the inherent risk; this is done by developing table of 'things to check', creating a formal record of what determines riskiness (note: this exercise also feeds directly into the statbarn classification, later). Ritchie (2014) suggests considering
    - Low frequencies
    - Extreme values, outliers and censoring
    - Lack of variation in the data
    - If the data is categorical
    - Dominance if the data are cardinal
    - The impact of absolute versus relative values
    - The likely presence of other information which could lead to uncovering detail (for example, for odds ratios it is assumed that marginal totals would be published in the paper, allowing the full 2x2 frequency table to be reconstructed from the odds ratio and some of the margins)
3. Identify any differencing risk, including hierarchies
4. Examine the statistic for other issues (for example, the root-index test for Herfindahl indices)
5. Identify whether any corrective measure effectively nullify the risk
6. Evaluate the risks, using the default-open approach; this should include
    a. likelihood of accidental production of risky versions of the stats
    b. the effort needed for deliberate attack eg brute force solution
    c. the incentives for deliberate attack
    d. the ease/reliability of corrective measures
7. Classify, including any corrective measures

While this process is clearly both subjective and inductive (it can't prove non-disclosiveness), it provides a mechanism for thinking about the problem and collating the evidence. We have reviewed new statistics in this manual using this approach.

The classification is based on the simplest form of the statistic: a top-N linear concentration ratio assuming that this is simply the sum of the largest N observations divided by the total, without any other adjustments; or a single-stage linear regression with no incidental parameters and all coefficient estimates being published. The assumption is that complexity reduces disclosure risk, but does not change its type: a complex unsafe statistic may be less unsafe but it still requires checking, while a safe statistic remains safe.

Finally we note that the classification holds in general for the statistic. However, ongoing work by the statistics team at UWE shows that there is general disclosure problem in all statistics: where a statistic is based upon small numbers and constructed values with a known finite range (for example, Likert scales), the values that generated the statistic may be determined exactly if there is sufficient precision in the statistic (Derrick et al, 2022a). This appears to be the case for a very large number of statistics; however, the likelihood of this drops off very rapidly with either an increase in observations, or more values for the variables. As such, this provides strong evidence for the use of thresholds greater than the statistically valid minimum of 3, but not practically effect the safe/unsafe discussion.

4. Classification of statistics (b): the statbarn model[1]

4.1 The statbarn concept

The statbarn model is a development of the safe/unsafe statistics idea. Analysts use a great range of statistical techniques. Devising statistical rules for all of these separately is not feasible. However, it is possible to combine statistics into groups based not on statistical relation but on common disclosure risks and solutions. For example:

- means and totals are identical for practical disclosure purposes
- a pie chart, a histogram or a scatter plot are all forms of frequency table
- multinomial logit, ordered probit, and proportional hazards all have negligible disclosure risk because of the non-linear convolution of variables

In some cases this requires additional assumptions. For example, the homogeneity between means and totals comes from the assumption that sample sizes are presented along with the means, a reasonable expectation. In a more complex case, an odds ratio *by itself* contains no disclosure risk *but* on the reasonable assumption that some of the marginal totals for control or treatment groups are presented along with the odds ratio, it is fair to assume the underlying 2x2 table can be reconstructed (Derrick et al., 2022b).

The *differences* between statistics are also important:

- means and frequencies share the risks of low numbers and potential for differencing
- but means have the potential for dominance
- survival tables are frequencies but generate an implicit secondary table

So a grouping would put means, totals, frequency tables and survival tables into three different disclosure groups (Figure 2):

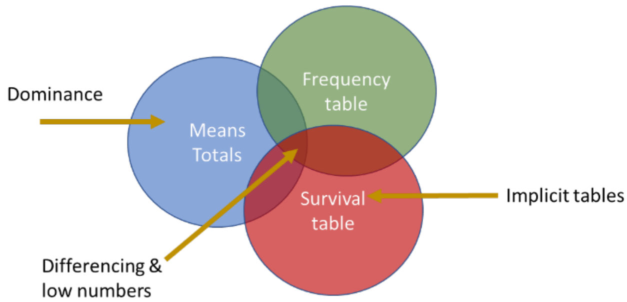


Figure 2 Differences between statistical types

Everything in the groups should have the same risks and solutions. For example, rounding or noise addition are valid solutions to disclosure concerns in both frequency and survival tables; but cell suppression is of limited value in survival tables because of a monotonic relationship between cells.

The advantages of this approach are both statistical and operational:

- Fewer rules/cases for researchers and output checkers to learn
- More consistent treatment of outputs

[1] This section text is largely adapted from Derrick et al, (2023), but the content in this section differs in several important ways from that conference paper: max/min now have their own category, smoothed distributions were moved into regression models, and the guidance on survival tables has changed significantly. These differences, reflecting barely two months additional analysis, demonstrate how this is still a developing field.

- Clearer distinctions between outputs
- Easier to develop the theoretical basis for any guidance
- Easier to update guidance when it changes (which it does)
- Output checker (and researcher) training can focus on the risky classes rather than trying to cover all cases

Checking statistics is now a case of 'what category does it fall into?' rather than 'what rules are needed?' Similarly, understanding the disclosure risk in a novel statistic becomes a question of classification. This is effectively defining a taxonomy of disclosure risk. Because classification, class and category is used in this field in many different ways, we refer to the groupings as 'statistical barns' or 'statbarns'[2].

The real value of this comes from finding that, in terms of disclosure characteristics, the minimum number of statbarns is fairly small. To a researcher, estimation of a hazard model bears little analytical relation to a quantile regression; but they pose the same disclosure risks: that is, no meaningful risk in any reasonable use, and so the only test needed is to make sure that this a genuine research use. In the case of estimated models, the tests are always

- Are there sufficient residual degrees of freedom (ie is this a model not an equation)?
- Is the model saturated; that is, the explanatory factors are all categorical and all fully interacted (ie is this a multi-way table masquerading as an estimate)?

And just like that, a large and essential part of research output is consigned to the disclosure risk box 'nothing to see here'.

Finally, this enables automatic output checking. The SACRO coders only need to know the statbarn code and then can draw all the information they require from a finite set of outputs. Ironically, it was the development of ACRO, a proof-of-concept for automated output checking developed for Eurostat and the precursor to SACRO (Green et al, 2021) that prompted the formalisation. Whilst the UWE team had been considering an improved taxonomy, efficient coding required that ACRO translated multiple Stata tabulation commands into a single command, providing a concrete example of the value of classification.

4.2 Constructing the statbarns

Each of the statbarns is designed to house statistics with

1. The same functional form
2. The same conceptual disclosure risk
3. The same conceptual SDC responses, even if in some cases these are not useful in practice (for example, cell suppression in survival tables)

Condition (1) may not be obvious, especially for derived values such as ratios. The key is to consider whether a statistic could be transformed into another, using the statistic itself and other reasonably available information. Hence, means and totals are equated in the group 'linear aggregations' because of the likely availability of the sample size. Simple ratios (such as linear top-N concentration ratios) are classed as linear aggregations, because (a) it is not unreasonable to suppose that the totals from which the shares are derived is presented along with shares, and (b) the shares themselves might be informative (eg '75% of the market controlled by the two largest firms').

[2] The concept behind this is of the farmer rounding up her animals into different groups: cows in the cowshed, chickens in the henhouse, goats in the trees and so on.

Other cases are not so clear. There is an argument for saying that, because the simpler statistical tests can be restructured as regression models with dummy variables, statistical tests and regressions should be lumped together. We have not explored this in detail, and we suspect that the saturation argument may be a difference between the two, and so we have kept those separate for the moment.

As with safe statistics, the assessment of functional form is based on the simplest version. A chain-linked index may have negligible disclosure risk but at the moment it is classified as 'index number', which puts it with 'frequencies'. These are areas where they may be future scope for efficiency gains by defining additional statbarns. The issue is likely to be whether a new class can be defined with sufficient clarity that the gains from more efficient rules are not outweighed by the extra time needed to assess which statbarn the output is in.

Because of condition (2), statbarns as a whole are classified as 'safe' or 'unsafe' statistics. There cannot be a mix, as this would imply differing disclosure risks. Condition (2) also helps to distinguish similar cases. For example, in the early stages of development, maxima and minima were included with 'position' statistics (median, percentiles, inter-quartile range etc). However, max and min present a class disclosure risk, as membership of the class is simply defined as membership of the sample; this is not the case for median, where membership of the class defined by the median requires knowledge of the data subject's ranking.

Finally, condition (3) provides some extra perspective for helping output checkers. For example, suppression is the only meaningful response to max and min: if an extreme value is informative, it seems unlikely that rounding or noise addition would meaningfully and effectively reduce that information content. For all of the statbarns classified as 'safe', there is generally no mitigation, as a disclosure risk from these implies a catastrophically poor piece of statistical analysis. The only exception to this is for regression, where a saturated regression can be re-assessed as a table (this is one reason why we may continue to distinguish between hypothesis tests and regression).

4.3 Current position

As it stands (October 2023), the SACRO models contains fourteen statbarns:

	Barn	Example	Class	Status
1	Frequencies	Frequency tables	Unsafe	Very well understood
2	Statistical hypothesis tests	t-stats, p-stats, f-stats	Safe	Provisional
3	Correlation coefficients	Regression coefficients	Safe	Confirmed
4	Position	Median, quartiles	Unsafe	Provisional
5	End points	Maximum, minimum	Unsafe	Very well understood
6	Shape	s.d., skewness, kurtosis	Safe	Provisional
7	Linear aggregations	Means, totals	Unsafe	Very well understood
8	Mode	Mode	Safe	Provisional
9	Non-linear concentration ratios	Herfindahl index	Safe	Provisional
10	Calculated ratios	Odds & risk ratios	Unsafe	Confirmed
11	Survival tables	Hazard/survival tables	Unsafe	Provisional
12	Gini coefficient	Gini coefficient	Safe	Provisional
13	Linked/multi-level tables	Nested categorical data	?	Unknown
14	Clusters	Cluster analysis	?	Unknown

Table 3 Statistical barns as at October 2023

It is clear that some of these statbarns cover a very large number of cases ('correlation coefficients' covers linear and non-linear regression, ANOVA, ANCOVA, pairwise correlation etc). In contrast, the disclosure risks of the mode are unlike any other statistic, and so it merits its own class. This shows the importance of identifying exactly what are the disclosure characteristics of a particular statistic.

The act of creating the list is itself a useful exercise, forcing one to consider what are the meaningful differences. For example, mean and median are often grouped together in OSDC guidelines, but they have quite different characteristics.

This list is likely to undergo change over time. Even in the development process, the list changed as more statistics were deemed to be of the same type, and others demanded a new type. The process of identifying risks and defining OSDC guidelines for each class is crucial, as this is usually the point at which it becomes clear whether a new type is needed or not. It may also be the case that trying to identify a minimal set is counter-productive. Initially maxima/minima were treated as special cases of percentiles; but in terms of communication of risk to researchers and output checkers, it was thought sensible to separate them. Finally, we have created some categories where, at the moment, we don't have enough information to be comfortable that they fit an existing category. Category 13 "linked/multiple tables" is an example – it seems like these should be covered by frequency tables, but we suspect there are nuances which need to be explored, and so creating it as a separate category show the need for more understanding. Similarly, we suspect that the decision trees used by operational researchers (but not the ones used by machine learners) may need a class of their own.

The coverage of OSDC theory is decidedly patchy. The 'status' column has four values:

- **Very well understood:** disclosure issues, things to be checked and protection mechanisms have been comprehensively studied and there is **a large literature** and a **consensus**
- **Confirmed:** these have not been so well studied (conclusions rest on **one or two papers**) but we are confident that the conclusions and guidance are robust, well-founded and comprehensive
- **Provisional:** the analysis is new (or radically different) in this project; we have confidence in our conclusions but this is based on extrapolation from other types, and from our own understanding; there is substantial further work to be done (for example, on the impact of extreme values on a statistic) before the classification can be confirmed; **the appendices to this document** outline current thinking
- **Unknown**: while we may have idea o the likely characteristics, basic analysis has not been carried out; these are therefore 'unsafe'

At present, the focus is to get the 'provisional' status raised to 'confirmed'.

The list above is provisional and was devised by the SACRO team based at the UWE. SACRO's network of output checkers was consulted as to whether this was a sensible approach in general; the response was positive and expected: earlier evidence-gathering sessions had already indicated a desire for simplification of the current OSDC landscape. The initial categories seemed both sensible and comprehensive, although these are likely to be modified as they develop in practice.

Of more concern to the output checkers was how they (and researchers) would easily check the guidelines for statistics. This is achieved by a look-up table, linking statistics to the appropriate statbarn, from which the corresponding checks, problems and solutions could be found. As the list is organised by disclosure characteristics, rather than use, this makes it harder for a new output checker (or a researcher) to discover the statbarn a statistic is in. We have created a second lookup table to facilitate this. The lookup table also has 'use type' – descriptive statistic, estimation, non-parametric modelling, discriminant analysis and so on. This reflects the way statistics are grouped together by researchers, and the way that statistics courses are taught. Table 3 summarises.

Statbarn	Descriptive statistics	Factors & discriminants	Hypothesis testing	Non-parametric statistics	Regression, estimation, modelling	Total
Calc. ratios		1		2		3
Clusters				1		1
Correlation					32	32
End points	3					3
Frequencies	7					7
Gini coefficient				1		1
Survival tables	1			1		2
Linear agg.	3				1	4
Multi-level tables	1					1
Mode	1					1
Position	3					3
Seq. analyses				1		1
Shape	3		1			4
Hypothesis tests		3	42			45
Grand Total	**22**	**4**	**43**	**6**	**33**	**110**

Table 4 Statistical barns by use type

This will be created as a searchable file, but the output tools being developed by the SACRO project (Smith et al., 2023) intend to incorporate this in the user front end. Researchers and output checkers should be able to click on a link to see more information about the output, drawn from the statbarn classification. In the initial project this will only include basic data such as that shown above, but in future it may be useful to expand the information on each classification.

4.4 Graphical outputs

Graphs do not present new issues. In theory, every graph can be represented as a table in some way, and so the above rules could be applied. To take an obvious example, a pie chart or a histogram are clearly just one-way tabulations, whereas a waterfall graph is a two-way table. As a counter example, a kernel density estimate could be represented as a mathematical form, but in practice is almost always show graphically. In practice, we need separate rules because (a) the quantity of information differs, and (b) precision is likely to be lower in a graph.

Consider the Kaplan-Meier graph, which is simply a survival table re-presented, usually in proportional form (percentage of initial cohort continuing). Survival tables are classed as 'unsafe but very low risk' because, even in the case of a unit being identified, the personal information content in the survival table is negligible. Griffiths et al. (2019) suggest that the underlying survival table should be supplied along with the graph, but this can cause more problems:

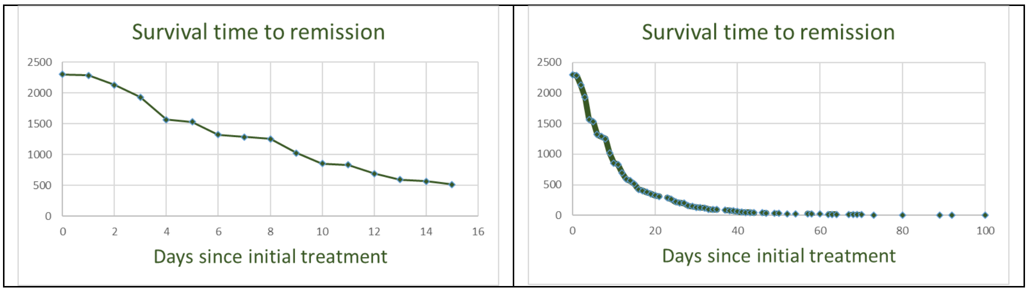


Figure 3 Example survival graphs

In the left-hand graph, the source table would have 15 steps and be checkable by a human. But that table would have precise numbers easily readable, whereas getting the exact figures from the graph depends on the way that the image was produced (and even then, some laborious analysis). In the right-hand diagram, identifying individual data points from the image has become harder whereas a survival table with 100 rows in it is much more likely to present a disclosure risk, as well as being harder to review.

The above graphs are presented as numbers. Formally Kaplan-Meier graphs should show the survival rate rather than numbers (ie 0%-100%). In theory this makes graphs slightly less disclosive than the survival table: tables reflect exact numbers, whereas the number of decimal points determines the accuracy of the graph points.

5. Operationalising SDC

5.1 Principles- vs rules-based OSDC[3]

All output checking starts from basic rules of what is and is not acceptable. How these rules are applied is the difference between the two approaches to managing output-checking for conformance to regulation: ‘rules-based’ and ‘principles-based’.

Under a rules-based approach (RBOSDC), a rule is a hard limit; no exceptions are allowed. Setting the rule is problematic as the rule is trying to balance both confidentiality and utility. Too restrictive a rule prevents the publication of useful but non-disclosive findings; too loose a rule allows many useful results to be published but increases the risk that disclosive results leak out.

Under the principles-based approach (PBOSDC), the rule becomes a ‘rule-of-thumb’: it guides decisions but is not always followed, and can be adjusted as necessary. The researcher can argue that the rule can be ignored if, and only if:

- the output is non-disclosive, and
- the detail in the output is important to the researcher, and
- this request for an exception is a rare occurrence for the researcher

The first condition is obvious. The second condition ensures that the output-checker and researcher only spend time negotiating over an output when the result matters to the researcher. This is appealing to researchers as it puts them in charge of deciding when something is ‘important’, rather than the output checker. The third condition ensures that researchers do not abuse the system. Note that the terms “important” and “rare” are not specified – this is an area for the researcher and output-checker to negotiate. As a result, training the researcher to understand the concept is necessary (see ONS, 2019, for example training).

PBOSDC is two-way: the output-checker can also argue that the rule-of-thumb is inappropriate in a specific case because it does not protect confidentiality. For example, the output-checker may argue that a higher threshold is needed because the data are particularly sensitive and the patients are easily identified. Some organisations (for example, National Records for Scotland) operate a two-tier system with a lower ‘regular’ threshold and a higher threshold for outputs based on more sensitive data.

PBOSDC systems can have stricter rules than RBOSDC: an overly restrictive rule-of-thumb does not limit research as the researchers always have the opportunity to argue for an exception. Hence, the threshold can be set high as it only has to address the confidentiality problem; the utility problem is dealt with by the exception mechanism.

Because rules-based is very limiting in research environments, most organisations claiming to be rules-based operate a ‘rules-based but sometimes…’ system allowing for ad hoc relaxation of rules. This can generate the worst of both worlds: inefficiency and uncertainty. However, PBOSDC has been around long enough now for it to become familiar. Most organisations seem to be moving towards PBOSDC in practice, if not in formally.

5.2 Output checking as a customer process

Alves and Ritchie (2019) argue that output checking is primarily an *operational* process, not a statistical one; that is, the effectiveness of the process is determined by understanding how inputs

[3] This sub-section is largely adapted from Alves and Ritchie (2021)

and outputs (in this case, statistical requests) flow through the system, rather than seeing the assessment itself as the aim of the process. Using models of customer segmentation from operations research, they reduce output checking to three different scenarios: the runner, the repeater, and the stranger (Table 1)

Type	Meaning	Output checker skills	Examples
Runner 90% of requests	Exactly the same service each time	None if output is comprehensible. Can be done automatically	Safe statistics Unsafe statistics that meet rules
Repeater 9% of requests	Similar services but needing some human intervention	Some understanding of statistics and willingness to make subjective judgments	Exception requests for unsafe statistics
Stranger 1% of requests	Novel services needing to be evaluated	Ability to assess statistics and define new rules (a statistic should never be a stranger twice)	New types of statistics

Table 5 Segmentation of output types and the role of process

This perspective highlights

- The SDC rules may have a statistical foundation but the application of those rules is primarily an operational decision (for example, choosing what to manage as an exception and how)
- Output checkers' skills can vary, depending on what they check, and may involve little or no skill (as for automatic checking) or extensive statistical knowledge
- Safe and unsafe statistics align directly with this model

As the FAQs (Part III) demonstrate, an effective and efficient output checking process requires attentive to procedural detail. Accordingly, in this guide we will consider operational issues when considering output checking.

5.3 Active Researcher Management

The co-operation of the researchers makes a big difference for the efficiency and security of output checking processes. The researcher is best placed to identify what outputs are important, carry out the initial assessment of disclosure risk, and make the appropriate corrections. The ease with which researcher outputs can be reviewed by output checkers depends on how well the researcher presents them. If the researcher has not produced high-quality outputs which must be rejected, or wants to ask for an exception under PBOSDC, the effectiveness of the conversations between the output checker and the researcher will be influenced by the attitudes of both parties to each other.

Desai and Ritchie (2009) referred to good practice in this area as 'active researcher management'. They noted that researchers are generally well disposed towards support staff, but are likely to be focused on their own goal (the production of research), and interference with or limitations on that goal may become a source of resentment. Similarly, user support officers who are overly prescriptive can create barriers to good working conditions.

This does *not* mean that support officers should not reject bad outputs or address inappropriate behaviours. It does mean that support officers need to recognise that effective communication relies on appreciating the goals, constraints and interests of researchers, on sharing information usefully, and on actively managing the understanding and expectations of the researchers. The role of output checker becomes less guardian and more pedagogue and facilitator.

Desai and Ritchie (2009) argue that active researcher management "…goes beyond setting up contracts and making sure researchers know their legal responsibilities. It requires making researchers partners in secure effective access – and *making sure that they understand this*” (Desai and Ritchie, 2009, p10 – emphasis added). In this guide, active researcher management is assumed.

Part II Guide to statistics

6. Review of statistical disclosure control solutions

This section provides a short summary of SDC solutions. For a more detailed review and discussions, see Green and Ritchie (2021).

The examples here are presented as tables, as these are where concerns are mostly likely to occur and the impact of alternative control methods can be seen. They are applicable to outputs in the following 'unsafe' statbarns:

- Frequencies
- Linear aggregates
- Position
- End points
- Calculated ratios
- Survival tables

Not all measures may be applicable in all cases – see the detailed discussion in Part 2 for relevance.

In the examples below, 10 is used as the threshold for disclosure.

6.1 Limiting source data combinations

Definition: Only permitting specific tabulations of variables which allow each cell to have sufficient observations to prevent disclosure risk. This often involves collapsing or merging categories to increase observation counts.

Example of table without output limiting:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39, 632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Same example with output limiting the classes to 'university' and 'other':

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
Other	39	£40,557	12	£39,937	51
Total	251	£40,773	116	£40,205	367

Table 6 Example of output limiting

This is a limit on what can be produced, rather than a check on what is produced. It is often used for Census data, or other cases where a fixed set of tables is to be produced. For research data, it is generally better to enforce such rules through limiting the set of variables available to the researcher, rather than relying on the researcher to check and enforce rules.

Pros

- Easy to demonstrate non-disclosiveness with a small number of variables and categories.
- Can be managed automatically table delivery systems

Cons

- Difficult to prove non-disclosiveness via differencing which increases with the number of potential categories and values.
- Does not allow for additional table combinations.
- Limited usefulness in research on underrepresented groups
- Where researchers have access to the microdata, very hard to ensure that researchers have used only the approved combinations

6.2 Hiding data

6.2.1 Cell suppression, with adjustment of totals

Definition: Removing cells which fall below threshold, usually replacing with blanks or some other non-informative filler. It requires additional suppression or recalculation of totals to prevent disclosure by subtraction. 'Suppression' is often used as a synonym for cell suppression, although formally this should be avoided as it could also refer to whole-table suppression.

Example table before cell suppression:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39,632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Example table after cell suppression (totals omitted):

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	<10	£39,632	-
Vocational School	18	£40,793	<10	£40,241	-
Total	251	£40,701	-	£40,115	-

Table 7 Example of cell suppression (totals omitted)

The row and column totals must be adjusted (or omitted) so that true values cannot be discerned by differencing. In the above example, the totals have been omitted. Row totals of (104, 355) and column totals of (21, 18, 355) would also be acceptable.

It is generally considered good practice to remove cells associated with the suppressed cells. For example, the mean average salaries associated with the suppressed cells present little disclosure risk, but they are based on low numbers and are at risk of differencing. Of course, if the minimum-threshold rule was applied to all cells in the table irrespective of the statistic being presented, this would eliminate those cells too.

Pros

- Cannot be unpicked
- Non-structural zeros can be handled in the same way as different undesirable values
- Easy to automate primary suppression
- Suppression is obvious to reader

Cons

- Information loss due to suppression
- Does not protect against disclosure by differencing
- Researchers may forget to adjust totals

6.2.2 Cell suppression, with secondary suppression

Definition: Removing cells which fall below threshold, along with other cells above the threshold to protect the hidden cells.

Example table before cell suppression:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39,632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Example table after cell suppression (with secondary suppression):

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	-	£40,320	-	£39,632	29
Vocational School		£40,793	-	£40,241	22
Total	251	£40,701	116	£40,115	367

Table 8 Cell suppression, with secondary suppression

This maintains marginal totals, although at the cost of missing data in specific cells. This is popular in official statistics, where consistency in totals across tables is valued, tables are generated in the same way from the same source, the set of tables is limited and predictable, and the organisations have the processes in place to ensure this is done effectively. We do not generally recommend this for research use, as only the first of the four conditions is likely to hold. Moreover, researchers are generally antagonistic to the removal of non-disclosive data.

Pros

- Cannot be unpicked
- Accurate totals are retained
- Non-structural zeros can be handled in the same way as different undesirable values
- Easy to automate primary suppression
- Suppression is obvious to reader

Cons

- High level of information loss due to secondary suppression as number of primary suppressions increases
- Secondary suppression removes values that are not in themselves a disclosure risk
- Does not protect against disclosure by differencing
- Researchers may struggle to do this effectively in large tables

6.3 Changing data

6.3.1 Noise addition – Simple

Definition: Tables are adjusted by adding a small amount of random noise so that the true value in the cell is uncertain.

Example table before noise addition:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39,632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Same example after noise addition:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	215	£40,974	107	£40,473	322
College	19	£40,331	8	£39,632	27
Vocational School	17	£40,798	6	£40,241	23
Total	251	£40,699	121	£40,115	372

Table 9 Simple noise addition

Noise can be additive (x->x+y where y is a random value -n to +n), scaled (x->x+my, where m grows as x increases) or multiplicative (x->xy).

The values associated with the cells may need to be adjusted as well. In the above example, these are means. Therefore, they could be left untouched, or separately have noise added. If the figures were totals, then changing the number of observations could change the interpretations of the totals considerably, and so some adjustment would seem essential.

Because noise addition is not obvious by looking at the table, it **must** be clearly highlighted in the methodological notes to the table. Some values will be unaffected by the noise, as occasionally adding zero noise should be one of the permissible outcomes for a random noise allocation process.

Pros

- Simple and easy to implement in automated table-production systems
- High degree of practical protection
- Less vulnerable to differencing than cell suppression when the table is generated statically (eg presented in a paper)
- Little impact on large values

Cons

- Small values are disproportionately affected if additive noise is used
- Larger cells can be significantly altered if noise is scaled to cell value, or multiplicative noise is applied
- Non-additivity of tables (ie different totals for the 'same' categories)
- The level of association between variables is affected
- Variance of cell counts is increased
- Viewers may mistakenly assume these are genuine values
- If the table is generated dynamically, repeated requests for the same table generate a differencing risk

6.3.2 Noise addition – Cell-key adjustment

Definition: Noise addition method that adds noise consistently across tables. A noise parameter is randomly assigned to every individual microdata record. When records are combined in cells, a deterministic function is applied to the combined noise values so that the same combination of cells always generate the same noise.

Pros and cons are the same as simple noise addition with the following additions/alterations:

Pros

- No risk from repeated requests of dynamic tables or differencing
- Particularly useful if the same information is being presented in different ways
- Can be applied automatically

Cons

- More complicated to understand, and to implement
- Sparse tables complicate specification of noise look up.

6.3.3 Noise addition – Differential Privacy

Definition: Method of noise addition that works to prevent disclosure by considering what values *could have* been in the dataset, not what actually were, and then adding noise. The "noise"-the random value that is added - ensures that no single person's inclusion or exclusion from the database can significantly affect the results of queries.

DP is easily automated and appears to provide a guarantee of security, so it is often used in the private sector (Google and Apple both use it). As a result, it has a higher profile than other methods.

Pros and cons are the same as simple noise addition with the following additions/alterations:

Pros

- For a single query, creates a mathematical guarantee of the probability of privacy to a level of probability set by the data holder

Cons

- Absurd results can be generated when distributions are highly skewed, or when rare events are being described
- Privacy guarantee only holds for single queries – multiple queries can create a disclosure risk
- Level of acceptable risk is subjective and may be set at a level that risks data exposure – it is not clear that the 'privacy budget' is well understood by users

6.3.4 Rounding – Conventional

Definition: Adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell.

Example table before rounding:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39, 632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Same table after rounding:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	210	£40,989	100	£40,473	310
College	20	£40,320	10	£39, 632	30
Vocational School	20	£40,793	10	£40,241	30
Total	250	£40,701	120	£40,115	370

Table 10 Simple rounding

The larger the base rounding value, the more protection is provided, although more accuracy is lost. For additional protection, rounding can be carried out to multiples of the rounding value eg

- Round to nearest 5 in 80% of cases
- Round to nearest 10 in 15% of cases
- Round to nearest 15 in 5% of cases

Again, there is a question of whether the values associated with the rounded cells should also be adjusted and/or rounded.

Pros

- Effective for protecting frequency tables, especially when one data set calls for many tables
- Protects small frequencies and zero values as it becomes difficult to identify genuine zeros
- Easy to implement
- Less vulnerable to differencing than cell suppression – it still exists, but requires work

Cons

- Totals and cells rounded separately may not add up, making unpicking feasible
- Increasing perturbation as base value increases – cells can be significantly altered by the rounding process and aggregation compounds these rounding differences
- Potentially risk of unpicking if there are many genuine zeros
- Inconsistencies in data may be visible to the reader and compromise researcher credibility.
- The level of association between variables is affected by rounding, and the variance of the cell counts is increased
- Small values are disproportionately affected, particularly if the base value is large

6.3.5 Rounding – Random

Definition: Rounding each cell to one of the two nearest base values in a random manner. Each cell value is rounded independently and has a greater probability of being rounded to the nearest multiple of the rounding base.

Pros and cons are the same as conventional rounding with the following additions/alterations:

Pros

- Random rounding is harder to unpick due to uncertainty around whether data was rounded up or down

Cons

- Increased loss of data accuracy

6.3.6 Rounding – Controlled rounding

Definition: Rounding using linear programming techniques to round cell values up or down by small amounts.

Pros and cons are the same as conventional rounding with the following additions/alterations:

Pros

- Additivity is maintained in the rounded table, making the table more realistic and harder to unpick.

Cons

- Computationally complex process that is not possible in some software.

6.4 Table redesign

Definition: Changing the categories used to display data.

Example of table:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
College	21	£40,320	8	£39, 632	29
Vocational School	18	£40,793	4	£40,241	22
Total	251	£40,701	116	£40,115	367

Same example with ‘college’ and ‘vocational’ categories merged:

Higher Education type	Male alumni	Male average salary	Female Alumni	Female average salary	Total Alumni
University	212	£40,989	104	£40,473	316
Other	39	£40,557	12	£39,937	51
Total	251	£40,773	116	£40,205	367

Table 11 Table redesign

Note that this example is identical to Table 3, output limits. The difference is that the decision on which groups to control is now under the control of the researcher, rather than being specified as a function of the data. Researcher training (eg ONS, 2019) advises the researcher to consider this in preference to suppression or perturbation. The reason is that suppression and perturbation relate to a particular table; redesign of categories forces the researcher to consider the structure of her data more broadly: in the above example, by reflecting on whether combining the two small categories makes sense. This increases the likelihood of the same changes being applied across multiple tables, reducing differencing risk.

Pros

- Retains accurate information
- Encourages the researcher to reflect systematically on categorical structures
- Clear to the user / reader
- Relationships between cells are not broken

- Most often used as a first choice for researchers

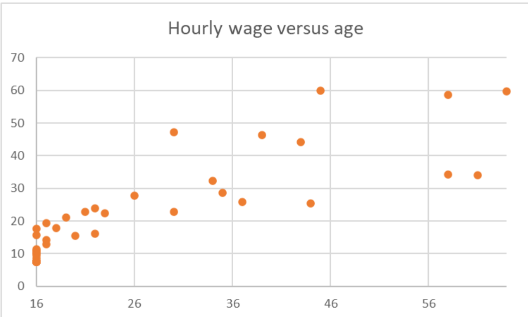
Cons

- Requires subject knowledge to apply
- Cannot easily be automated
- Loss of precision
- Relationships between cells may be weakened

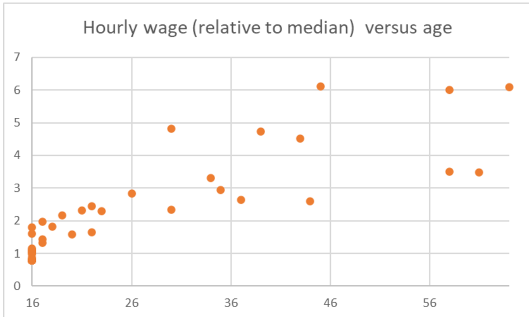
6.5 Data transformations

When the data is transformed, disclosure risk is decreased. Whether this reduces risk sufficiently depends on the transformation. In general, linear transformations make a small difference to disclosure risk, whereas non-linear ones make a much larger one.

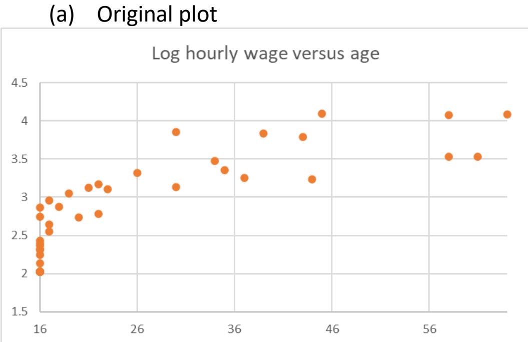
Consider these three scatter plots of hourly wage versus age:



(a) Original plot



(b) Converted to relative to median



(c) Converted to natural logarithms

Figure 4 Alternative transformations of a scatter plot

The (linear) transformation from absolute to relative wages makes little meaningful difference to the disclosure risk, especially if the base value – in this case, the median hourly wage – is published. In contrast the log graph compresses large values and stretches the gap between small values; large outliers are less noticeable, and the axis no longer shows easily-interpretable values.

Whether this formally reduces risk depends on the display medium. If these were .SVG files with the point values encoded in the data, then in theory neither transformation makes any difference because both can be converted back to the original value – in fact the median transformation is safer as it requires more work on the part of an attacker. In practical terms however, the log transformation is more likely to dissuade any de-transformation attempt.

6.6 Reducing precision

The argument in the previous section, that formally the transformation makes no difference, depends upon the data points being measurable with a large degree of precision. If however calculations are done and only reported to a few decimal places, with those shortened values being graphed, then the data is protected after the transformation. Reducing precision is a valid way to reduce risk in certain cases.

Consider the following table which show proportions with varying degrees of precision:

	N	Proportion				Implied counts		
		0 dp	1 dp	2 dp				
English	8200	80%	80.4%	80.39%		8160	8201	8200
Scottish	1100	11%	10.8%	10.78%		1122	1102	1100
Welsh	600	6%	5.9%	5.88%		612	602	600
N. Irish	300	3%	2.9%	2.94%		306	296	300

With a large base value, 2 decimal places are necessary for the proportions to reflect actual numbers. As the number of observations shrinks, so does the number of decimal places: with 10% of the observations, one decimal place exactly reproduces the source numbers; with 400 observations instead of 10,000, just the simple percentage is enough to identify the counts.

In the case of the graphs displayed above, reducing precision and then graphing the results would have had much more impact on the log values, which are much smaller and more closely spaced than the relative wage values.

So, reducing precision is a valid disclosure control tool, but the context needs to be considered.

7. Guide to classified outputs

7.1 How to read this section

Each statistic can be assigned to a statistical barn based on typology and structure. This section details all of the 'statistical barns' ('statbarns', the classes of statistics) identified and analysed o far. The table below summarises how this section should be used.

- Researchers wanting to prepare an output for clearance should read parts 1-6
- Researchers wanting an exception to the rules should read part 7 to see if an exception is likely to be requested, and what information an output checker will need
- Human output checkers should read section 8 for background
- Automatic output checkers are based on the guidance in section 5 only

1	Summary	This is a very brief reference for those familiar with the statbarn.
2	Description of statbarn	Examples of the statbarn. For a more detailed list, see the spreadsheet in the project web page, which also provides a detailed list of types and the barns into which they fall
3	Risk factors	The elements that create problems for this statbarn.
4	Classification	Classification as safe or unsafe; within the latter, classified as low-high risk, depending on whether this is a realistic risk or not.
5	Criteria for rules-based approval	Rules to be applied for automatic approval. Note that there can be different rules for machine-based approval.
6	Remedial action	List of remedial actions to correct problems that either will have been applied, or what should be suggested to researcher to apply.
7	Issues to consider if an exception is requested	The factors that should be taken into account if an exception is requested under a principles-based regime. This only considers how to assess disclosiveness. We assume that the other factors (such as importance, rarity of the request) have already been considered. We take the perspective outlined in Part 1 on how to assess disclosure: assume that the output is non-disclosive, consider the circumstances which could make it disclosive, and assess their reasonableness.
8	Underlying theory/discussion	Further detail (how well are the risks understood, stability of thinking etc), plus references. This section is not necessary for researchers.

7.2 Frequencies

7.2.1 Summary

Examples of type	Frequency tables, histograms, shares
Safe or unsafe?	Unsafe
Risk level	High
Risk elements	Low numbers Differencing Complementarity Class disclosure Undisclosed linked units Categories
Checks to be made	Thresholds, categories
Appropriate responses	Suppression, rounding, noise addition
Covered in automatic tools	SACRO (threshold and statbarns) tArgus sdcTable

Modelling	Very well understood
Key text(s)	Hundepool et al. (2010) Griffiths et al. (2019) Most SDC textbooks and guides

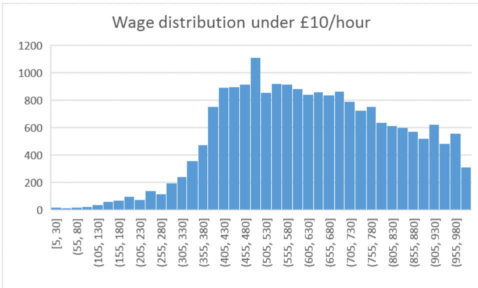
7.2.2 Description of statbarn

This statbarn covers frequencies i.e. counts of things, either in tables (most common), in certain graphs such as histograms or bar charts, or single as in a description of the number of survey participants.

Examples:

Age last birthday	Below NMW	At or above NMW	Total
16	0	1366	1366
17	0	1258	1258
18	114	990	1104
19	63	1003	1066
Total	177	4617	4794

(a) Frequency table



(b) Histogram

" The 25-29 year-old sample consisted of 36 men but only 7 women"

"There were no examples of males aged 61-65 presenting"

(c) Text description

Table 12 Examples of frequencies being displayed

The frequencies can also be represented as percentages. It is important to treat these as equivalent to the underlying numbers, as we assume that the totals are available somewhere else in the research publication.

7.2.3 Risk factors

Low numbers	Single observations may allow an individual to be identified or have values attributed to her/him Two observations may allow one of the individuals to make inferences about the other Three observations is deemed safe on the assumption of no collusion Most organisations require a threshold higher than three to provide extra confidence If an **evidential zero** (ie something you would normally expect to be zero) is not zero, this should be investigated; similarly for evidential 100%
Differencing	Two tables, with N and N+1 respondents, generate an implicit table where the characteristics of the extra respondent are exposed
Complementarity	Binary categories have a complement which may not be included in the table eg a cell '83% white' generates an implicit cell '17% non-white'; this is a special case of differencing
Zero or full cells (class disclosure)	Cells with no entry in a category, or all the entries for a category, may lead to class disclosure. In Table 1a above, it is sufficient to know that an employee is 16 or 17 to know that they are paid at or above the minimum wage; it does not matter that there are 2,624 individuals with this characteristic. An **evidential or structural zero or 100%** (ie something you would expect) can be ignored. In Table 1a above, the data are from 2002, when there was no minimum wage for 16-17 year-olds; we therefore expect these cells to be zero.

Multiple units	A cell may appear to contain multiple units, but they may all be under the control of one survey respondent and so should be treated as one observation. Examples: four GP surgeries all part of the same Clinical Commissioning Group; five Census respondents identified as part of the same family
Category choice	The category choice may itself reveal information. For example, if a histogram for male wages goes from £4.50 to £10, but the corresponding histogram for females goes £4.00 to £10, this implies only females are in the category £4-£4.50

7.2.4 Classification

This is an **unsafe** statistic. Within unsafe statistics, it is **high risk**.

7.2.5 Criteria for rules-based approval

For manual checking:

a. Apply threshold set by your organisation (note that this may vary by dataset)
b. Check for structural zeros or full cells
c. Check that the underlying unit is genuinely independent (for example, that the seven people in the group are not members of the same family, or all the references are to students at one school)
d. Check that the categories are comprehensive and apply to all data – in particular, check for complements

For automatic output checking:

a. Apply threshold set by your organisation
b. Check for zeros or full cells
c. Check, as far as possible, that the categories are comprehensive and apply to all data

7.2.6 Remedial action

All of the remedial actions suggested in Part 2 Section 6 are appropriate. Cell suppression and table redesign are most common. Note that totals **must** be adjusted if suppression, noise addition or rounding are used (or that **secondary suppression or controlled rounding** is applied).

7.2.7 Issues to consider if an exception is requested

Factors required to make the output potentially specifically identifying individuals, or potentially creating a class disclosure

- The categories used on the table can be reasonably matched up with 'public' characteristics of an individual (ie things such as height or holiday destination, not private as such e.g. appointment time or salary)
- Either
  - Inclusion in the dataset is known or reasonably assumed, and
  - The categories relate to very few people in the dataset
  - *Example: population is all taxpayers in Wales; category is submarine engineers*
- Or
  - The combination of categories is so unusual as to identify individuals in the population
  - *Example: senior ranks in the clergy, by gender*

- Or (for class disclosure)
    - Inclusion in the dataset is known or reasonably assumed, and
    - Individuals can be unambiguously put into one class

It is also worth considering whether the above could apply to a differenced table:

- Is there a reasonable possibility of differencing to uncover an individual or group?

If you are confident there is no reasonable risk of identification (**real or perceived**), approve the exception.

If the identification can be **reasonably** made, consider

- Is the identification provided by this output likely to be linkable to other outputs to reveal the characteristics of the identified individual
- Is the revealed information new (it is not structural eg "Bob is in the dataset, as he is an employee working in Bristol and the dataset is tax data from every male working in Bristol")?
- Is the revealed information sensitive (ie even if this is not technically sharing unknown information, this is a revelation of information that it would be unethical to highlight)?

If the identification can reasonably be made and this reveals something informative about the person, do not release; otherwise consider granting the exception. Remember to consider that membership of the dataset itself may be an issue; for example, "Wilberforce is in the dataset; this is a survey of swingers; therefore Wilberforce is a swinger".

You may also want to consider whether the release of the output, even if non-disclosive, generates a **perception** of identification. Just one observation in a cell may lead to false positives, which could have reputational risks for the organisation.

7.2.8 Underlying theory/discussion

SDC of frequency tables has been discussed for decades. The theory is simple, the problems are obvious and there is no real debate about the topic. Any SDC textbook will cover disclosure control of frequency tables in extensive detail, and say much the same thing as any other, which is summarised above.

The choice of a threshold is arbitrary; a background to threshold choice can be read in Ritchie (2022). From a theoretical perspective a threshold of 3 is the minimum to protect against statistical disclosure; in practice we see a variety of different thresholds being applied. The threshold applied reflects the organisation's risk preferences, and may vary across datasets as well. The threshold can have a signalling function: choosing a higher threshold (such as a minimum of 100 observations in each cell) for some datasets indicates that these datasets are to be more 'protected', even if the practical effect on risk exposure is negligible.

The only debate is whether differencing risk is a genuine problem in real research environments. It clearly is a theoretical risk but there is very little evidence to support or refute this. In official statistics environments there is some evidence (see Smith et al. 2012). However, this arises because official statistics are, by design, meant to be consistent alternative cuts of the same underlying data. In contrast, researchers are likely to cut and re-present the data in multiple ways without clear explanation; there is ample of the non-reproducibility of research results, even from a known dataset. As a result, it is likely that differencing in a research environment, is likely to be a problem only within a project and not across projects.

With no definitive evidence one way or the other, controlled environments prefer to reduce the likelihood of differencing being a problem by increasing the threshold above the statistical minimum of three so that **likelihood** of differencing is reduced (while the differencing between tables with 3 and 4 observations and 23 and 24 observations is still one, a gap of exactly one is less likely when the minimum number of observations is higher). As thresholds are usually set above this level anyway, this is a case of killing two birds with one stone.

The exception to this is where the differencing arises because of complementarity generating an implicit table T → 1-T. Humans and computers do this very differently. Humans are better at identifying where complementarity is an issue, computers are comprehensive but inefficient. Consider two tables and what each sees:

N=500 Source table		Human checks…		Computer checks…	
English	73%	Not English	27%	Not English	27%
English	73%	Not necessary		Not English	27%
Scottish	12%			Not Scottish	88%
Welsh	7%			Not Welsh	93%
N. Irish	4%			Not N. Irish	96%
Other	4%			Not Other	96%

Table 13 Humans versus computers assessing complementarity

A human recognises that all the options are covered in the second case. The computer just recognises a series of binary variables and calculates the complements, unaware that this is unnecessary. However, unless the dataset is very large, the time penalty is likely to be negligible.

7.3 Statistical hypothesis tests (SHTs)

7.3.1 Summary

Examples of type	T-test, chi-square, R^2, standard errors
Safe or unsafe?	Safe
Risk level	n/a
Risk elements	Insufficient degrees of freedom
Checks to be made	Residual (not model) degrees of freedom
Appropriate responses	n/a
Covered in automatic tools	SACRO
Modelling	This document
Key text(s)	None – see below

7.3.2 Description of statbarn

This covers all the statistical hypothesis tests ie that test "Is X greater than/equal to/less than/different to Y". This includes tests generated as part of a wider complex output (R^2, F test, z-scores generated during a regression run), standalone tests ("is the mean of this group the same as the mean of that group?", or structural tests ("is the distribution of animal-related deaths the same across all social classes?").

Examples:

```
      Source |       SS           df       MS      Number of obs   =       808
-------------+----------------------------------   F(4, 803)       =     14.77
       Model |  1222.94161         4  305.735403   Prob > F        =    0.0000
    Residual |  16625.6779       803  20.7044556   R-squared       =    0.0685
-------------+----------------------------------   Adj R-squared   =    0.0639
       Total |  17848.6195       807  22.1172484   Root MSE        =    4.5502

------------------------------------------------------------------------------
     surplus_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
sh_inc_activity |   .4104541   .5970432     0.69   0.492    -.7614955    1.582404
  sh_inc_grants |  -1.490512   .5899031    -2.53   0.012    -2.648446   -.3325779
            big |  -.6748409   .3678892    -1.83   0.067    -1.396979    .0472971
 sh_staff_costs |  -3.787034    .692105    -5.47   0.000    -5.145582   -2.428485
          _cons |   3.990169   .4097552     9.74   0.000     3.185851    4.794486

Ramsey RESET test using powers of the fitted values of surplus_pc
       Ho:  model has no omitted variables
                 F(3, 800) =     10.71
                  Prob > F =      0.0000
```

(a) Tests generated by regression model

```
Hausman test:  Ho:  difference in
coefficients not systematic
                  chi2(3) =        6.50
                Prob>chi2 =      0.0897
```

(b) Standalone test

Table 14 Examples of statistical hypothesis tests (shaded boxes)

In the above examples, the SHTs are found in the red boxes.

Note that degrees of freedom may be misleading: for standalone tests, the *model* degrees of freedom may be the only ones reported – 3 in the case of the Hausman chi-square. The important number is the **residual degrees of freedom**: N- model restrictions imposed. For most practical purposes, the number of residual degrees of freedom approximates to N.

7.3.3 Risk factors

Low residual degrees of freedom	Common parametric tests (t-test, higher order ANOVAs) are equivalent to regressions and we include this for completeness.

7.3.4 Classification

This is a **safe** statistic.

7.3.5 Criteria for rules-based approval

For manual and automatic checking:

a. Check minimum degrees of freedom and approve

7.3.6 Remedial action

Not relevant

7.3.7 Issues to consider if an exception is requested

Not relevant

7.3.8 Underlying theory/discussion

Some SHTs can be rewritten as regression models by appropriate use of dummy variables, and so these could be considered as part of this category. However, SHTs have lower risk.

All SHTs are some variant on sum$((x-f(y))^2)/z$, where z and y may be functions of x or constants. This convolution prevents extraction of information by direct inspection or differencing. In addition, the SHT is a single number encapsulating all of the model information. We can therefore assume that, if the test can be carried out, the test statistic is safe.

Potentially one could consider very small numbers eg one could run a non-trivial t-test on just three observations, and if the original values were limited (eg integers in the range 0-10), and the test statistic was reported at sufficient decimal points, then possibly the original values could be recovered. But this is such an extreme case we can ignore. In any case, doing a dof check to most threshold (5 or 10) stops this. Hence even the most wilfully incompetent researcher is not going to release information through a chi-square, say.

We will be reviewing this provisional classification in 2024.

7.4 Coefficients of association

7.4.1 Summary

Examples of type	Estimated coefficients, models, AN(C)OVA, correlation tables
Safe or unsafe?	Safe
Risk level	n/a
Risk elements	Insufficient degrees of freedom Saturation in category-only models
Checks to be made	Degrees of freedom, saturation
Appropriate responses	n/a
Covered in automatic tools	SACRO
Modelling	Well understood
Key text(s)	Ritchie (2019)

7.4.2 Description of statbarn

This statbarn covers coefficients derived from estimating statistical models. It includes linear and non-linear estimation, ANOVA, ANCOVA, and pairwise correlation coefficients.

Examples:

Regression analyses of determinants of surplus

```
      Source |       SS           df       MS      Number of obs   =       808
-------------+----------------------------------   F(4, 803)       =     14.77
       Model |  1222.94161         4  305.735403   Prob > F        =    0.0000
    Residual |  16625.6779       803  20.7044556   R-squared       =    0.0685
-------------+----------------------------------   Adj R-squared   =    0.0639
       Total |  17848.6195       807  22.1172484   Root MSE        =    4.5502

------------------------------------------------------------------------------
   surplus_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sh_inc_activity |   .4104541   .5970432     0.69   0.492    -.7614955    1.582404
  sh_inc_grants |  -1.490512   .5899031    -2.53   0.012    -2.648446   -.3325779
            big |  -.6748409   .3678892    -1.83   0.067    -1.396979    .0472971
 sh_staff_costs |  -3.787034    .692105    -5.47   0.000    -5.145582   -2.428485
          _cons |   3.990169   .4097552     9.74   0.000     3.185851    4.794486
```

Table 15 Examples of coefficients of association (highlighted)

7.4.3 Risk factors

Low degrees of freedom	If the model is estimated with K parameters and N observations, and N is only slightly larger than K, then the correlation approximates to an equation; full knowledge of the explanatory variables would allow one to exactly specify the dependent variable.
Saturated model	If the model is completely specified in categorical variables, with **all possible interactions** as explanatory variables, the regression should be viewed as a multi-dimensional table showing the mean of the dependent variable in the cells. A special case of this is regression on a single binary variable as the only explanatory variable.

7.4.4 Classification

This is a **safe** statistic.

7.4.5 Criteria for rules-based approval

For manual checking:

b. Check that the **residual** degrees of freedom (number of observations less number of variables and any other restrictions – in the above example, 803 'residual df') exceed your organisational threshold value
c. Check that the model is not saturated ie not all variables are categorical and **fully** interacted see discussion -below
d. If the model is saturated, it can be assessed as a multi-dimensional table; however, this is now a magnitude table and should be assessed as such (in particular, you now need to ask for counts in each category as these can't normally be determined from regression output)
e. Reject if a regression with a single binary explanatory variable (should have been submitted as a table)

For automatic output checking:

a. Check that the residual degrees of freedom (calculated by model) exceed minimum
b. Reject if a regression with a single binary explanatory variable

7.4.6 Remedial action

None. If the model fails the degrees of freedom check, it is not a model. If it fails the saturation check, it should be described and assessed as a magnitude table.

7.4.7 Issues to consider if an exception is requested

There are no meaningful exceptions.

7.4.8 Underlying theory/discussion

This is comprehensively covered in Ritchie (2019). Earlier versions of the paper provide slightly different perspectives on it, but perceptions have also changed since the earliest paper on this (Ritchie, 2006); the recommendation therein, repeated in Brandt et al (2010), to hide one of the coefficients, is no longer considered necessary.

The principle is straightforward. Coefficients of association are the result of convoluted calculations where everything depends on everything else. This is too complex to be unpicked, and cannot be differenced because of the interdependencies.

Theoretically, an estimate could be constructed to reveal a value, but this requires a set of highly implausible actions and extreme data conditions, and so can be ignored. It is also possible that a researcher could choose to deliberately engineer the estimate to release a specific value. This makes no sense in an RDC environment, where all data is visible to the researcher. It makes marginally more sense in a remote job server, where the researcher cannot see the base data, but all RJSs monitor and record all code, and it would be clear that the code has been manipulated as there is no legitimate reason for the necessary transformations. Therefore, deliberate manipulation of results can be ruled out.

Even in the worst cases – of the model failing the checks but this not being noticed – the downside is limited.

In the case of insufficient degrees of freedom, consider N=K+1 ie no degrees of freedom at all once the model process is taken into account. This is an exact equation. However, this does not directly reveal values; it only does this if someone has access to the explanatory variables, in which case the dependent variable for those observations can be perfectly predicted. This is bad, but we are still not saying that the coefficients themselves are directly revealing.

In the case of a saturated model not being noticed, the coefficients, in the appropriate combinations, reveal the mean value of the dependent variable for the relevant combination of characteristics. For example, this example is taking from Run 1 of the Output Checking Course exam:

Dependent variable: hourpay

Variable	model1	model2	model3
male	-2.7746 ***	-2.8796 ***	-2.8660 ***
white	0.0807	0.0193	0.0178
white_male		0.1169	0.1046
age			0.7852 ***
age_sq			-0.0089 ***
_cons	11.0601 ***	11.1154 ***	-4.7170 ***

N.obs: 31,410

Table 16 Example of a saturated regression (model 2)

In this example, the researcher runs three wage regressions with explanatory variables:

- Model 1: male, white dummies
- Model 2: male, white dummies; interaction between dummies
- Model 3: male, white dummies; interaction between dummies ; scale variables

Model 2 is the fully saturated model. It effectively reproduces the table showing average wages for four groups:

	Male	Not male
White	8.1382	11.1347
Not white	8.2358	11.1154

Table 17 implied table underlying the saturated regression

This is a magnitude table and so has the standard problems of magnitude tables described above – low numbers and dominance (but not class disclosure). Again, the coefficients themselves are not necessarily revealing but might be in certain circumstances. Note that in the exam, the vast majority of candidates failed to notice the saturation.

What happens if the model is genuine, but there is a single non-zero observation for one variable (for example, the 'other' option in a set of categories)? In this case, the error on that observation would be zero, and the dependent variable could be predicted exactly – if, **and only if**, all the explanatory variables were known. Given that this also requires the unique observation to be known, this is a very high information requirement, and can be reasonably ignored.

It should be re-iterated that even being close to the limit on degrees of freedom or complete saturation would require the researcher to be carrying out some spectacularly bad statistics. Even allowing for the less competent researcher, the likelihood of either limiting case occurring in a genuine research environment is too small to be worth considering.

Finally, these results above only hold for the most simple models. Anything that requires, for example, two-step estimation (as in robust estimation) or incidental parameters (as in longitudinal or clustering models) is completely non-disclosive. Hence, there is no need to consider more complex models.

7.5 Position

7.5.1 Summary

Examples of type	Median, percentiles, maxima, minima
Safe or unsafe?	Unsafe
Risk level	Low
Risk elements	Class disclosure
Checks to be made	Threshold counts
Appropriate responses	Suppression
Covered in automatic tools	SACRO
Modelling	Reasonably well understood
Key text(s)	This document Brandt et al (2010) Griffiths et al. (2019)

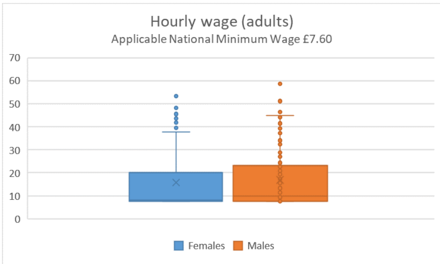
7.5.2 Description of statbarn

Considering ordering all the observations for a particular variable from smallest to largest. This statbarn considers what can be known from highlighting a particular point on the scale.

Examples:

	Females	Males
Min	7.60	7.60
Bottom 10%	7.60	7.60
Median	8.22	9.79
Top 10%	36.83	35.58
Max	53.30	59.93
Number of obs	80	132

(a) Distribution table



(b) Box plot with whiskers

Table 18 Examples of position variables (highlighted in distribution table)

7.5.3 Risk factors

Class disclosure	A percentile value may reveal something about the individuals in that group. In table (a) above, every male in the bottom half of the distribution earns below £9.79

7.5.4 Classification

This is an **unsafe** statistic. Within unsafe statistics, it is **low risk**.

7.5.5 Criteria for rules-based approval

For manual checking:

a. do the numbers for that group (and complementary groups) meet the threshold?
    i. **Do not** use the number of observations **at** the class boundary

In example (a) above, for males there are 13 people in the top and bottom deciles, but for females only 8 people in the top and bottom 10%. On a threshold of ten, the male figures would be approved but the female figures would fail. The median is fine for both.

For automatic output checking:

d. Carry out threshold check on percentiles identified, including complements

7.5.6 Remedial action

Suppression, rounding and noise addition are appropriate.

7.5.7 Issues to consider if an exception is requested

Factors required to make the output potentially disclosive

- The percentile boundary (including min/max at 0%/100%) being displayed is informative about the people in that group, or
- There is an informative class disclosure because upper and lower bounds of a group are the same

and

- Membership of the group is reasonably known (is there other information that definitely puts an individual at that point of the distribution?)

Unless both of these hold, grant the exception.

7.5.8 Underlying theory/discussion

In general, locating someone on a distribution is not easy. Percentile boundaries themselves do not in general say what happens within those boundaries, and so only contain limited information. Therefore, we would generally consider these low risk and unlikely to be disclosive.

There is sometimes confusion between thresholds **within** a class and thresholds **at** the class boundary. The first of these is correct. Suppose the median wage in a dataset is £11.90 per hour, and the threshold is 5 people. It does not matter if five or fewer people earn exactly £11.90 per hour; what matters is that there are least five earning **at or below** £11.90, and at least five people earning **at or above** £11.90. It does not matter whether there are less than five people *at* the median.

7.6 Extreme values

7.6.1 Summary

Examples of type	Maxima, minima
Safe or unsafe?	Unsafe
Risk level	High
Risk elements	Small numbers at the value Class disclosure
Checks to be made	Threshold counts Class disclosure
Appropriate responses	Suppression, rounding, noise
Covered in automatic tools	SACRO
Modelling	Very well understood
Key text(s)	This document Brandt et al (2010) Griffiths et al. (2019) Most SDC guides

7.6.2 Description of statbarn

This statbarn considers what can be known from presenting the largest or smallest value of a variable. Examples:

	Females	Males
Min	7.60	7.60
Bottom 10%	7.60	7.60
Median	8.22	9.79
Top 10%	36.83	35.58
Max	53.30	59.93
Number of obs	80	132

Note: minimum wage in 2018 is £7.60 for adults

(a) Distribution table

"Wages rates varied from £7.60 to £59.93 per hour"

(b) Text description

Table 19 Examples of max/min (Highlighted in distribution table)

7.6.3 Risk factors

Identification at extreme points	Maxima or minima may refer to a single individual. In example (a) above, the maximum hourly wages are likely to refer to specific individuals
Class disclosure	If membership of the dataset is known, extreme values reveal something about members. In the above example, **every** male in the dataset earns £59.93 per hour or less.

7.6.4 Classification

This is an **unsafe** statistic. Within unsafe statistics, it is **high risk**.

7.6.5 Criteria for rules-based approval

For manual checking, the recommended approach is:

a. Is the maximum or minimum **informative** or **structural** (for example, 0% or 100% when the value is a proportion)?
b. If informative
    i. does it meet the threshold at the max min? **and**
    ii. is the value informative about all members in the dataset?

In example (a) above, the maxima are informative (telling you something about people) and so should be checked but the minima are structural (describing the feasible structure of the dataset – the minimum possible wage) and can be ignored. The maxima for males and females are likely to represent a single person and so will not meet the threshold. However, see the 'Underlying discussion' below; there are operational reasons why banning max/min unless an exception is requested may be preferred.

For automatic output checking:

a. Block and only allow as an exception (but see discussion below)

Note that an alternative (more user-friendly but theoretically more risky) is to carry out a threshold check, which will automatically block unique maxima/minima, and may identify structural zeros. Again, see below for operational reasons.

7.6.6 Remedial action

Suppression is appropriate. Rounding and noise addition are possible but unlikely to be successful. If an end point is enough of an outlier to be informative, blurring its value is not going to have much effect.

7.6.7 Issues to consider if an exception is requested

Factors required to make the output potentially disclosive

- The min/max is informative about people at the extreme who can be reasonably identified, **or**
- The max/min is informative and membership of the group is reasonably known

Unless one of these holds, grant the exception.

An **informative** extreme value is one where it would not reasonably be expected. In the above example, the lowest wage at the legal minimum is not informative. The maximum wage (translating to about £120,000 per year) may be informative about teachers, but not informative about senior lawyers. An **uninformative** value could be a limited one eg study was terminated after 20 weeks so records are right-censored.

7.6.8 Underlying theory/discussion

Maxima and minima are class disclosures as they relate to everyone in the dataset. Membership of the dataset is sufficient to show that these values are associated with individuals; hence in the past the default behaviour is to block them.

However, humans can take a reasonable judgement about whether the extreme value is informative, and so for manual checking, blocking all max/min unless an exception is requested may be over-protective. However, the no max/min rule unless an exception is requested can be simpler to explain.

For automatic checking, the computer does not know whether a max/min is structural or not. Blocking all max/min unless an exception is requested is therefore a clear solution.

However, it is also possible for an automatic checker to ascertain the number of records at the min/max. If the number of records exceeds the relevant frequency threshold, then (a) this no longer refers to an individual and (b) this is evidence, albeit not conclusive, that the min/max is structural. Of course, this does not deal with the class disclosure issue if membership of the dataset can be inferred. Organisations may want to consider which approach is preferred.

Note that in the longer term it may be possible to infer other automatic rules eg if all values are non-negative, than this may indicates a minimum of zero is structural.

7.7 Shape

7.7.1 Summary

Examples of type	Standard deviation, skewness, kurtosis
Safe or unsafe?	Safe
Risk level	n/a
Risk elements	Residual degrees of freedom, differencing
Checks to be made	Degrees of freedom
Appropriate responses	No mitigation
Covered in automatic tools	No
Modelling	None
Key text(s)	Appendix to this document

7.7.2 Description of statbarn

This covers standard deviation, skewness and kurtosis, and higher orders if anyone really wants them.

Examples:

Weekly wage for those working in public or private sector, all adults, Jan 2004

	N	Mean	Std. deviation	Skewness	Kurtosis
does not apply	15,241	-9	0	.	.
no answer	138	135.5517	179.003	1.583294	6.118772
private	35,923	222.4125	292.5825	3.364227	40.06159
public	13,761	288.4439	252.4099	1.16707	5.715295

Table 20 Example of shape variables (highlighted)

7.7.3 Risk factors

Low degrees of freedom	If the statistic is estimated with N observations (N<=5), then the correlation approximates to an equation; the underlying variable values could be determined
No variation in SD	If the standard deviation is zero, this implies the underlying variable has no variation
Differencing	If the statistic is rerun with one additional observation, the additional observation can be identified without access to the mean, given sufficient degrees of freedom

7.7.4 Classification

This is a **safe** statistic.

7.7.5 Criteria for rules-based approval

For manual and automatic checking:

a. Reject if statistic is standard deviation and is exactly zero
b. Otherwise, approve if N> degrees-of-freedom threshold

Note that formally we should be checking if N-K is above the degree-of-freedom threshold (where K=1, 2 or 3 for SD, skewness kurtosis respectively). N>d.o.f. threshold is a reasonable approximation and much simpler.

7.7.6 Remedial action

There is no meaningful mitigation if there are insufficient observations.

7.7.7 Issues to consider if an exception is requested

Not relevant

7.7.8 Underlying theory/discussion

Shape statistics involve the sums of deviations raised to powers of two or more. These cannot be directly unpicked. However, because these are univariate, there is a potential differencing risk. Two SDs, that differ in a single observation could lead to the disclosure of that observation, simply by knowing the two SDs and one of the means (not both). For this risk to materialise, the researcher has to publish the two SDs and one mean with one observation difference, at sufficient decimal places to allow the value to be uncovered exactly (remembering that the new observation would change both the mean and the variation around it). This is possible, but not sufficiently likely to exercise concern in practical environments. See Appendix 2 for a discussion.

The only concern is whether the statistic is zero. In the case of the standard deviation, this would imply that the variable being studied only has one value $x_i = \bar{x}\,\forall i$, effectively a class disclosure. It might not be important but we ban it. If the user wants to report no variation, they can report directly rather than using derived variables.

7.8 Linear aggregates (means and totals)

7.8.1 Summary

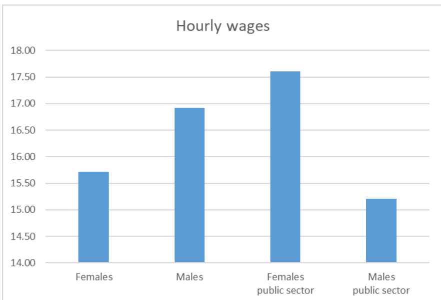
Examples of type	Means, totals, simple indexes, linear concentration ratios
Safe or unsafe?	Unsafe
Risk level	High
Risk elements	Low numbers Differencing Dominance
Checks to be made	Thresholds, dominance
Appropriate responses	Suppression, rounding, noise addition
Covered in automatic tools	SACRO (threshold and dominance) tArgus sdcTable
Modelling	Very well understood
Key text(s)	Hundepool et al (2010) Griffiths et al. (2019) Any SDC text

7.8.2 Description of statbarn

This statbarn covers sums, totals, simple indexes, linear concentration ratios (ie sums of the shares in a total – see "non-linear concentration ratios" below for an example) and other linear aggregations. These may come in tables (most common), in certain graphs such as bar charts, or single as in a description of survey characteristics.

Examples:

		Female	Male
All employees	N	150	150
	Wage	£15.72	£16.92
Public sector	N	69	64
	Wage	£17.61	£15.21



" Mean time to treatment was 17.3 days"

"In total, 679 ambulance dispatches were made in the survey period"

(a) Magnitude table (b) Bar chart (c) Text description

Table 21 Examples of mean and totals being displayed (highlighted in (a) and (c))

7.8.3 Risk factors

Low numbers	Single observations may allow an individual to have values attributed to her/him Two observations may allow one of the individuals to make inferences about the other
Differencing	Two tables, with N and N+1 respondents, generate an implicit table where the values for the extra respondent are exposed (directly for totals, with some maths when means are displayed)
Dominance	The data may be so distorted by one or two outliers such that the values for those individuals may be estimated with a reasonable degree of accuracy by someone looking at the statistic; the wages of a Premiership footballer living in a small village may be guessable to some accuracy by looking at the average village wage,

7.8.4 Classification

This is an **unsafe** statistic. Within unsafe statistics, it is **medium risk**.

7.8.5 Criteria for rules-based approval

For manual and automatic checking:

a. Apply threshold set by your organisation (note that this may vary by dataset)
b. Check P-ratio and N-K dominance rules if possible

In manual output checking, it may not be possible unless the researcher produces the necessary figures. In that case, the output checker needs to use their own judgement as to whether dominance is likely. To assess this consider:

- Are there few observations?
- Are there some observations which are several orders of magnitude greater than others?

If the answer to either of these is 'no', dominance is not likely to be an issue. For the rationale behind this, see the discussion below.

The dominance rules are as follows. In both cases, assume the variable of interest is ordered such that x[1]>x[2]>…x[N], and that there are N observations in the data. The parameters p%, k% and T will be set by the data owner.

Rule	Definition	How to calculate it	Rationale
P-ratio	Pass if $\frac{(\sum_{i=3-N} x[i])}{x[1]} > p\%$	Sum the smallest values x[3]…x[N] Divide by the largest value x[1] If this is greater than p% there is no dominance on this rule	The purpose is to prevent the supplier of the second largest value, x[2], combining that information and a reasonable guess about the size of x[3]..x[N] to estimate the value of the largest unit x[1] to less than p%
N-K	Pass if $\frac{(\sum_{i=1-T} x[i])}{(\sum_{i=1-N} x[i])} < k\%$	Add up the T largest values Divide by the total If this is less than k%, there is no dominance on this rule	The aim is to ensure that the largest T observations do not effectively count for the whole of the statistic, so that we should treat it as just relating to that observation. Note that the conventional name for this is the N-K rule even this makes for a confusing nomenclature. WE have used T in the example to represent the top T observations.

Consider this example with 20 data points, ordered from largest to smallest.

£2,301	£624	£171	£49	£16	£7	£5	£4	£4	£4	£4	£4	£4	£4	£4	£4	£4	£4	£4	£4

If p% = 10%, k% = 90 and T = 2, then the p-ratio is 12.9% and the N-K ratio is 90.1%. Therefore this set of observations would pass the p-ratio test but fail the N-K ratio test.

7.8.6 Remedial action

All of the remedial actions suggested in Part 2 Section 6 are appropriate.

7.8.7 Issues to consider if exception is requested

Factors required to make the output potentially disclosive

- There is a reasonable possibility that a contributor could be identified (see criterion for this under 'frequencies')
- If there is no dominance (the exception is requested for low numbers):
    - There is a reasonable likelihood of differencing from similar statistics being produced
- If there is dominance:
    - It is likely that the dominant observations would be known, and
    - The information revealed is useful and novel

7.8.8 Underlying theory/discussion

As for frequencies, SDC of magnitude tables is a well-trodden path. Any SDC textbook will cover much the same material as is summarised above.

While dominance exercises the theoreticians, it rarely has a practical effect. It is not at all obvious even to the output checker when there is dominance in the data. Moreover, the conditions for it require extreme values (orders of magnitude greater than most observations).

The p-ratio is a high bar; even the highly skewed example given above passes this test. The p-ratio also requires that the second largest contributor has a very good idea of the value of observations x[3]..x[N], as well as its place in the pecking order, both of which are difficult assumptions to justify normally. If the second-largest unit is very small, it can be confident of the values of the units x[3]..x[N] as its own value acts as a cap, but not of its own position – is it really second-largest? If it is clearly the second-largest, it is largely to find it hard to gauge the values of the other smaller units.

The N-K rule is easier to fail, particularly in business where a couple of units might dominate a market. For example, turnover in the UK oil industry may fail the N-K rule because BP and Shell are much larger than any competitors. On the other hand, dominance does not mean that a value is identified, only that an approximation to it is deemed 'close enough'. It is likely that BP and Shell know far more about each other's operations than can be usefully inferred from a simple statistic.

In manual output checking, it may not be able to check for dominance without the necessary figures being provided by the researcher who has the necessary data to calculate them. In these cases, the output checker's knowledge of the data is invaluable. Without doing a formal check, an output checker can be confident that dominance will not be a problem unless we have all three of

- The data cover a small geographical area, a small industrial sector or some other sector where membership of the group is easily established
- The largest couple of entries are each as large as the rest of the observations put together
- The largest entries are easily identifiable (famous person in a small village, one dominant business)

In general, though, output checkers should not be spending much time on dominance checks without good reason.

7.9 Mode

7.9.1 Summary

Examples of type	Mode
Variations	None

Safe or unsafe?	Safe (assuming not all same value)
Risk level	Low
Risk elements	All values same
Checks to be made	All values same
Appropriate responses	None
Covered in automatic tools	SACRO [TBD}
Modelling	Well understood
Key text(s)	This document

7.9.2 Description of statbarn

This statbarn covers the mode i.e. something reported as the most common value.

Examples:

"Modal height is 178cm"	"The most common eye colour is brown"

Table 22 Examples of mode

7.9.3 Risk factors

Values the same	All observations having the same value means that the mode is effectively a class disclosure

7.9.4 Classification

This is a **safe** statistic, as long as not all values are the same.

7.9.5 Criteria for rules-based approval

For manual checking:

a. Check mode is not only value – use any evidence that other values exist (eg mean, median or percentiles are different), or check with user

For automatic output checking:

a. Order observations by size x[1]>x[2]…>x[N]
b. Check x[1] ≠ x[N]

7.9.6 Remedial action

There are no remedial actions. If there is only one value for all observations, this counts as a class disclosure.

7.9.7 Issues to consider if exception is requested

Factors required to make the output potentially disclosive (note that this implies we have already discovered that all observations share the mode)

- The modal value is surprising and interesting

If this is not the case, grant the exemption.

7.9.8 Underlying theory/discussion

Brandt et al (2010) and subsequent references argue that the mode is safe because, by definition, it is the most common value; therefore

- If everyone observation has a unique value, the mode is a random pick amongst these; it has no value other than the confirm the existence of at least one occurrence of that value ("modal weekly wage is £272.37")

- If several observations share the mode, there is no information how many observations that represents

Only if all observations share the same value does it become a class disclosure. If there are missing values, then this condition is not met and the mode is safe.

More recently the question of differencing was raised. Suppose that the following situation occurs:

- In a dataset, 50 individuals have green eyes and 50 have blue eyes. An additional observation is included, with blue eyes.

Initially, both green and blue are valid modal values; it is not determined. In the second case, the mode is blue. If the mode was initially reported as green, reporting the revised mode would reveal that the additional observed person has blue eyes (if the initial report of the mode was blue, then the additional observation is uninformative about the extra individual).

While this is a possibility, it requires several factors to come together:

- Two potential modes A and B, each equally likely
- A single additional observation with one of the potential modes, A
- Reporting of the mode twice, with the initial mode being reported as B

Other alternatives can be constructed, around definite changing to indefinite mode, and taking away rather than adding one observation. These cases are not impossible, but they seem highly unlikely in genuine research cases. To deal with this would require an additional rule to check whether there is a difference of one or zero observations between the reported mode and the next best candidate for the mode. This is not a function implemented in statistical programmes. Even if a single case was revealed, the nature of the mode, as the most common value, also suggests it is the least interesting value. Overall, we conclude that the carrying out additional checks on the mode (beyond all values being the same) is not easily doable, and is not likely to generate an appreciable risk reduction.

7.10 Non-linear concentration ratios

7.10.1 Summary

Examples of type	Herfindahl-Hirchsmann index, diversity index
Safe or unsafe?	Safe
Risk level	n/a
Risk elements	Dominance of largest value
Checks to be made	N, index value
Appropriate responses	None
Covered in automatic tools	SACRO?
Modelling	None
Key text(s)	Appendix to this document

7.10.2 Description of statbarn

This statbarn covers non-linear concentration ratios, where the ratio is not just a simple sum of the shares. Simple concentration ratios, such as ‘market share of the top 5 firms”, are covered in linear aggregates.

Example:

Values	62	45	17	12	3	2	2	1	1

Shares	42.76%	31.03%	11.72%	8.28%	2.07%	1.38%	1.38%	0.69%	0.69%
Squared shares	18.28%	9.63%	1.37%	0.68%	0.04%	0.02%	0.02%	0.00%	0.00%

Simple concentration ratio (sum of top 3): 85.52%   Herfindahl index (sum of squared shares): 30.06%

Table 23 Example of linear and non-linear concentration ratio (highlighted)

A ratio such as the top-3 ratio above is a linear combination, and is treated the same as any other total. The Herfindahl index is non-linear in its value (ie a change in a source variable value does not always lead to the same proportional change in the statistic).

7.10.3 Risk factors

Dominance	If the largest value is close to 100%, then the square root of the index approximates to it

7.10.4 Classification

This is a **safe** statistic but with a specific criterion attached.

7.10.5 Criteria for rules-based approval

For manual and automatic checking:

a. Check N>2
b. Check H< some threshold value (eg 0.81), which reflects a minimum level of uncertainty about the share of the largest value

7.10.6 Remedial action

No mitigation

7.10.7 Issues to consider if an exception is requested

Even H is greater than the threshold value, the researcher can make the case that the values are uninformative. For example, in Green et al (2019), a Herfindahl index is used to assess the dependence on four funding sources of 150 different charities. The index in this case was safe as there was no indication which of the funding sources dominated for any charity, or which charity any figure related to.

7.10.8 Underlying theory/discussion

Generally safe statistics do not require the enforcement of rules; they are understood to have negligible risk in all plausible research uses. However, Ritchie (2007) does note that if simple unambiguous rules can be applied easily and consistently to confirm there is negligible risk irrespective of data, then these can be classified as safe. This is the case here. N>2 ensures no collaboration, and the same values can be given by different combinations of variables; in a worst case of 3, normalised to a total of 100:

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
obs 1	45	58	80	81	92	92
obs 2	45	22	20	12	8	4
obs 3	10	20	0	7	0	4
H	0.42	0.42	0.68	0.68	0.85	0.85
$\sqrt{H}$ Overestimation of x_1	-43%	-12%	-3%	-1%	0%	0%

Table 24 Herfindahl index

The index is no reliable guide to any values other than the largest even in this simplest case.

The largest value is an issue. The accuracy of the root of H as an approximation to the largest value approaches 100% as the largest share approaches 100%. Once the largest share gets above 90%, this is largely independent of the value of any other observations. Hence, it makes sense to check that the largest value is sufficiently different from the root of H; or alternatively that the largest does not exceed, say 90% (almost equivalently, that H does not exceed 0.81). These are checks which need to be made but once confirmed they are sufficient to guarantee non-disclosure.

Why does this count as a ‘safe’ statistic? Because, given the checks, we can be confident there is no disclosure. It is possible that the root of the index reflects the largest value but unless the largest share is very large, other distributions will give the same result. Moreover, this is a ratio, and so without the total it is of limited disclosure risk. Finally, we note that there is no meaningful differencing risk, as an additional observation changes the total and so all the shares as well.

Hence we note that the disclosure risk is uniquely located in either (a) just 2 observations, so a risk of one unit finding out about the other, and we ensure that N>3; or (b) a threshold for an acceptable value – which will also cover the case of N=1 as H=1 in this case.

7.11 Odds ratios, risk ratios and other proportionate risks

7.11.1 Summary

Examples of type	Calculated odds ratio, risk ratio NB NOT estimated risks from eg proportional hazards models
Safe or unsafe?	Unsafe
Risk level	Low
Risk elements	Marginal totals published elsewhere
Checks to be made	Frequency checks on source table
Appropriate responses	Rounding
Covered in automatic tools	No
Modelling	Minimal
Key text(s)	Derrick et al (2022b) Appendix in this document

7.11.2 Description of statbarn

The odds and relative risk ratios reflect the likelihood of a particular outcome between treatment and control groups.

Examples:

Contingency table		Diseased	Healthy	
	Smoker	82	231	313
	Non-smoker	7	625	632
		89	856	945
Odds ratio	31.69			
Relative risk	23.65			

Table 25 Odds and relative risk ratios (highlighted)

7.11.3 Risk factors

Marginal totals published elsewhere	The ratios themselves pose no disclosure risk. However, it is likely that the ratios combined with other information could expose frequencies

7.11.4 Classification

This is an **unsafe** statistic, but **low risk**.

7.11.5 Criteria for rules-based approval

For manual and automatic checking:

a. Require the underlying contingency table to be produced; check as for any frequency table

7.11.6 Remedial action

Rounding the ratios may be sufficient, especially if there are many observations.

7.11.7 Issues to consider if an exception is requested

As for frequency tables.

7.11.8 Underlying theory/discussion

The ratios themselves are not disclosive, and nor can differencing expose them. However, we make the assumption that some of the frequencies, or some marginal totals (row or column totals in the contingency tables above) are published in the paper. When combined with the ratio, all the values in the table can be reconstructed in a wide variety of cases (Derrick et al. 2022, provide examples). Hence, approve the contingency table, not derived statistics.

These are low risk because one would not normally construct these ratios when there are very very few cases, so thresholds in real datasets are likely to be comfortably met. Even if the output checker forgot to check the contingency table, the exposure risk is limited.

7.12 Hazard and survival tables

7.12.1 Summary

Examples of type	Tables of survival/death rates, Kaplan-Meier graphs
Safe or unsafe?	Unsafe
Risk level	Low risk
Risk elements	Total number of observations; absolute dates; differencing; categories
Checks to be made	Dates and number of observations
Appropriate responses	Combining dates right censoring
Covered in automatic tools	Not yet
Modelling	Some in older guides, but no longer valid
Key text(s)	Appendix to this document

7.12.2 Description of statbarn

Tables of survival rates and/or death/exit rates from a starting position. Kaplan-Meier is a normalised graph derived from these.

Examples:

Day	Surviving	Deaths	Survival rate	Death rate	Hazard rate
0	2300	(original population)			
1	2286	14	99%	1%	1%
2	2131	155	93%	7%	7%
3	1930	201	84%	16%	9%
4	1565	365	68%	32%	19%
5	1532	33	67%	33%	2%
6	1322	210	57%	43%	14%
7	1287	35	56%	44%	3%
8	1255	32	55%	45%	2%
9	1023	232	44%	56%	18%
10	854	169	37%	63%	17%
11	834	20	36%	64%	2%
12	690	144	30%	70%	17%
13	591	99	26%	74%	14%
14	564	27	25%	75%	5%
15	512	52	22%	78%	9%

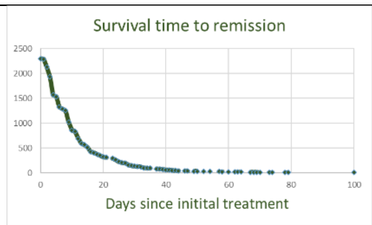


Table 26 Survival table and graph

7.12.3 Risk factors

Number of observations	The table shows N_s individuals with the same characteristics. By the end of the period, N_e individuals survive. For consistency with frequency tables, N_e should meet the frequency threshold
Absolute dates	If exit dates are absolute (Jan 1st, Jan 2nd etc) these pose an identifying risk and should not be associated with individual exits
Differencing	If the full dataset shows a single exit at a single time, multiple cuts of the dataset allow the characteristics of that person to be identified
Class disclosure (maximum times)	Extremely long exit times (outliers) may disclose individuals

7.12.4 Classification

This is a **unsafe** statistic, but it is low risk

7.12.5 Criteria for rules-based approval

For manual checking:

a. Confirm N_e meets usually frequency threshold T
b. Confirm exit dates are relative, not absolute
c. Confirm high values are not informative, or the data is right-censored (eg study finishes at an exit time to six weeks with some patients still in treatment)
d. If repeated versions of the table are presented with different subsets, confirm there are no single exits at point in time in the complete dataset

For automatic output checking:

a. Approve if no dates with single exit, **and**
b. Right censoring ie N_e>1

7.12.6 Remedial action

For absolute dates, convert to relative dates.

If absolute dates reveal single observations, combine dates. **Avoid suppression if possible** as it is difficult to get right (figures in survival tables can be calculated in multiple ways).

For outliers, use right-censoring to ensure that N_e>1.

7.12.7 Issues to consider if an exception is requested

The condition $N_e > T$ can be ignored if $N_s > T$ is greater than the threshold. This is for consistency with frequency tables to ensure this isn't a way to get out numbers below the threshold.

If absolute dates are used and there is a single observation at an exit point, approve if it is unlikely that the individual would be identifiable (including self-identification). If there are no single observations, absolute dates are fine.

Right-censoring is not necessary if all participants leave the study by an uninformative time (eg if median treatment time is 3 weeks and mean is 4 weeks, a last exit at eight weeks does not seem exceptional).

7.12.8 Underlying theory/discussion

Previous opinions on this have drawn analogies with frequency tables, and have therefore sought to block single exits in time periods. This ignores that **everyone** in the survival table **has the same characteristics** except exit date. Therefore it is the table as a whole that needs the frequency check. The table rows (exit dates) contain no identifying information, and so a single observation does not uniquely identify any member of the data being tabulated. Even to self-identify, a participant would have to be sure of both entry and exit dates; this differs from frequency tables where we assume that the participants' category values (age, gender etc) are simple and reliably known.

This does change in a small way if we have absolute dates – in that case, there is a slightly higher risk of confirming one's membership in the dataset, although again this does not contain any new information **about** members. To prevent the perception of disclosure, we do not allow absolute exit dates unless and exception is requested; but as this is a very limited risk we should normally allow this exception.

There is the potential for outliers. Right-censoring resolves this. Again, because individuals are not distinguished within the table except by exit date, any right censoring which has at least two people at the censor point adds sufficient uncertainty to mask any specific case.

Finally, consider someone producing multiple cuts of the data to identify characteristics. The full data shows one person left on day 7. Tabulation by gender reveals this person is male; tabulation by age group shows this person is not in the 25-34 age group; tabulation by ethnicity show this person is in the Chinese category; and so on. In this way it is possible to build up the characteristics of the single person who exit on day 7.

To do this requires that there is **a unique observation on the superset of information.** If the full table showed 2 people leaving on day 7, none of the subsets reveal unambiguous information. If the new tabulation is not a subset, it does not reveal information. A tabulation by gender shows one male leaving on day seven; a tabulation by male and ethnicity shows one Chinese man leaving that day; but this could be read directly for the second table. Moreover, the second table shows that one Chinese man left on day seven, but other Chinese man left on other days, so the person is not identified. And if the second table was not a subset (ethnicity, but not gender) then it cannot reveal the gender of the single person leaving on day seven.

This differencing holds true for other frequency tables; why do we need to consider it here? Simply because, in frequency tables we do not generally allow single observations in categories as they are uniquely identified as the only example of that combination of category. In survival tables, we allow single observations as the observable categories must have enough indistinguishable people in them to allow the table to be generated. But that does raise the (very small) possibility of this kind of differencing. If you are faced with many different cuts of the data, it is valid to challenge these are a potential disclosure risk; **but** the risk only arises from subsetting the data, hence we only need to confirm that there are no single observations in the broadest classification.

7.13 Linked or multi-level tables

To be done. Low priority

7.14 Cluster analysis

To be done. Low priority

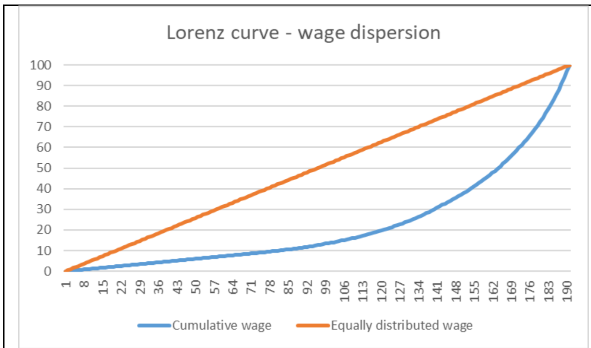
7.15 Gini coefficients

7.15.1 Summary

Examples of type	Gini coefficient
Safe or unsafe?	Safe
Risk level	n/a
Risk elements	2 observations
Checks to be made	N>2, coefficient <100%
Appropriate responses	None
Covered in automatic tools	Not yet
Modelling	None
Key text(s)	This document

7.15.2 Description of statbarn

Examples:



Gini coefficient = 55%

Table 27 Lorenz curve and associated Gini coefficient

7.15.3 Risk factors

Two observations	In this case, relative values can be calculated from the ratio
Coefficient=1	Either just one observation, or all observations have the same value

7.15.4 Classification

This is a **safe** statistic.

7.15.5 Criteria for rules-based approval

For manual and automatic checking:

a. Allow if N>2, **and**
b. The coefficient is less than 100%

If N=2 then you can work out the size of each share. Beyond that, not doable. If N=1, obviously the coefficient is 1. If N>1 but the coefficient is 1 (100%), all values are the same.

7.15.6 Remedial action

n/a

7.15.7 Issues to consider if an exception is requested

n/a

7.15.8 Underlying theory/discussion

The Gini coefficient is a measure of inequality from 0 (perfect equality) to 1 (perfect inequality). The discrete formula is (according to Wikipedia)

$$G = \frac{2\sum_i i x_i}{n\sum_i x_i} - \frac{n+1}{n}$$

This is potential identifiable if you have n=2 and one of the values, but otherwise, there seems little to worry about. It can't be differenced, because a new observation would alter the ranking, unless the new observation was larger than any other in the dataset (not entirely silly, as you might want to see the effect of adding a very rich person to a mix, for example in teaching). But even then I think you would need to do a lot of manipulation with uncertain results.

Fairly confident this is correct. Will check calculations later

8. Graphical outputs

The section will be developed further after consultation with user groups.

All graphs can be associated with one of the statbarns. It might be that some graph types require a new statbarn (for example Q-Q probability plots), but for now we assume we can fit all into the existing taxonomy. Currently we have allocated the following:

Graph type	Statbarn	Classification
Alluvial flow	Frequencies	Unsafe
AUC/specificity curves	Not classified yet	Unsafe
Bar graph	Linear aggregations	Unsafe
Box plot	Position	Unsafe
Cluster analysis (dendrogram)	Clusters	Unsafe
Density plot	Correlation coefficients	Safe
Heat map	Frequencies	Unsafe
Histogram	Frequencies	Unsafe
Kaplan-Meier	Hazard/survival tables	Unsafe
Kernel density plot	Correlation coefficients	Safe
Line graphs	Frequencies	Unsafe
Lorenz curve	Gini coefficient	Safe
Mean plots	Linear aggregations	Unsafe
Pie chart	Frequencies	Unsafe
Q-Q probability plots	Not classified yet	Unsafe
Scatter graph	Frequencies	Unsafe
Scatter plots	Frequencies	Unsafe
Smoothed Histogram	Frequencies	Unsafe
Waterfall chart	Frequencies	Unsafe

Table 28 Graphs and statbarns

In theory, the same rules apply to graphs as to their numeric equivalents. However, there are some differences in the assessments of graphical output:

- some graphs are 'exceptions' by their nature eg scatter plots are effectively sparse two-way tables with a lot of counts of one
- graphs may be more or less precise than their tabular equivalents; for example, low resolution graphs may have less detail from tables; but SVG graphics files contain the exact values to be displayed, which is probably more detailed than would normally be presented in a table (and there are tools on the web to extract data even from transformed images such as those in PDFs)
- extracting values from a graphic image requires more effort than reading them from a table
- graphs typically present much more information than a table, and so the volume of information can provide protection
- outliers may be more easy to see on a graph than in a table

Overall, it is likely that graphical output present less of a risk than tabular outputs, due to effort required to extract data. However, this is a consideration when considering an exception for a graph: initially the standard statbarn rules should be applied (eg does a pie chat have the minimum number of observations for each segment?).

Part III Support, advice and references

9. Output checking processes: the Decision Tree of Doom

The DRAGoN output checkers training course uses the ‘Decision Tree of Doom’ to help provide output checkers through the process. The full decision tree is

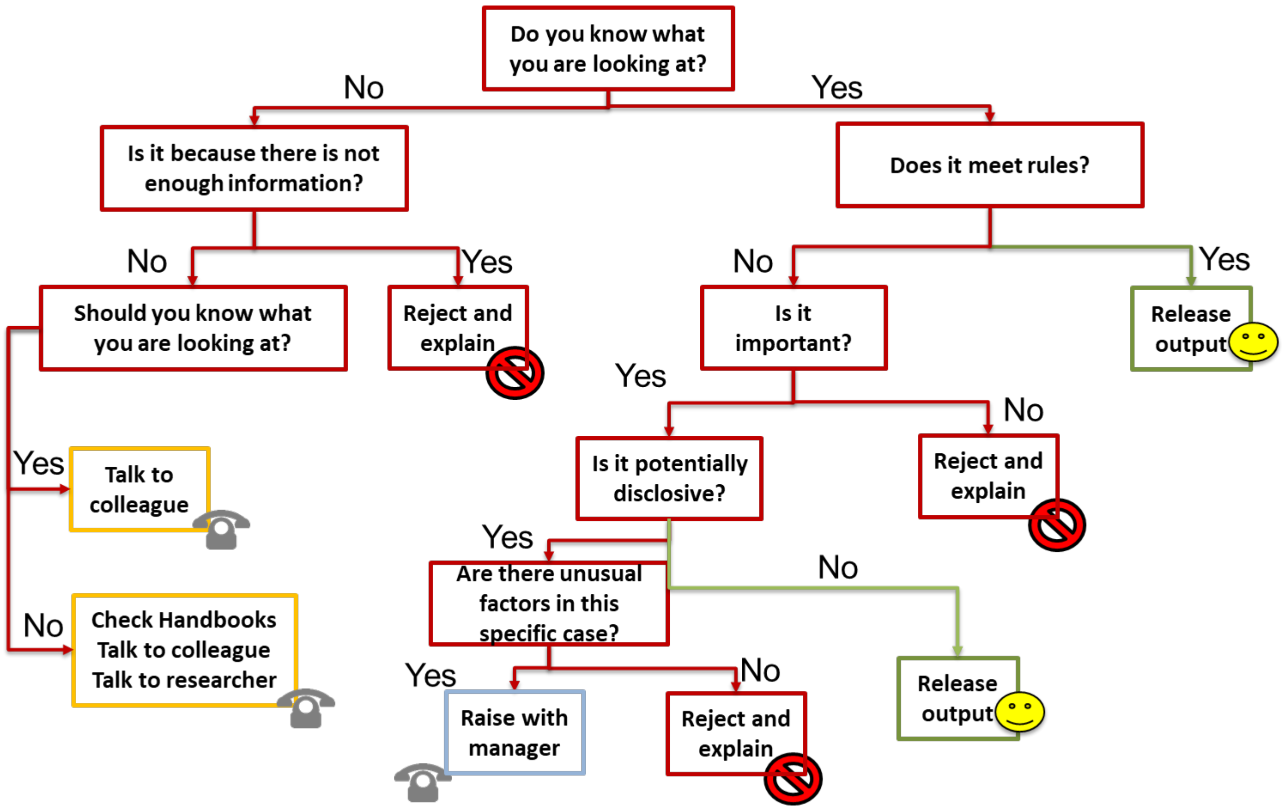


Figure 5 The Decision Tree of Doom

We take each part in sections.

9.1 Do you know what you are looking at?

Consider this path:

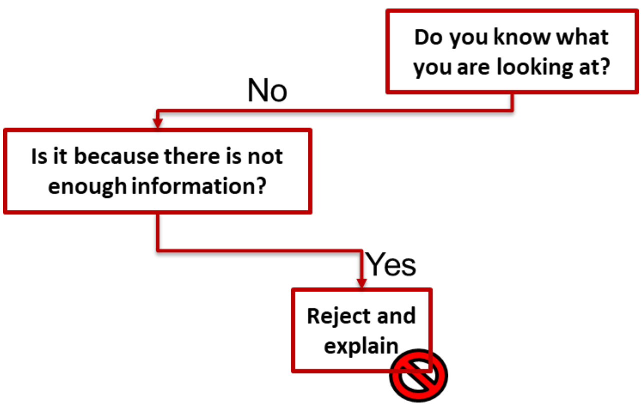


Figure 6 DToD: rejection for lack of information

This is straightforward. If the researcher has not provided enough information to assess the output for disclosure risk, reject it – ideally also explaining to the researcher what the problem is, so that the researcher can learn.

If the researcher seems to have provided enough information but the output checker does not recognise, the question becomes: is this something I should be expected to know about? Checkers are unlikely to be familiar with all types of output, but

- colleagues may have relevant knowledge
- handbooks may be able to help
- if this looks like something unexpected, researchers are generally very ready to talk about their research

So the path leads to acquiring more information:

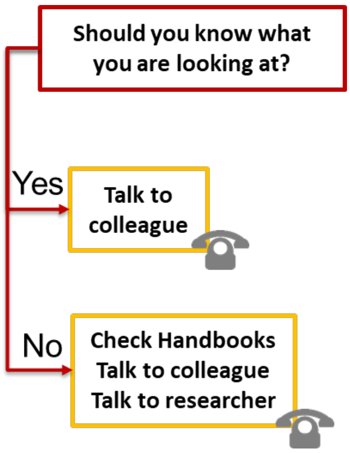


Figure 7 The DToD: gaps in output checker knowledge

9.2 Does it meet rules?

Once the output checker has ascertained what she is looking at, the next stage is to assess it against rules or rules-of-thumb:

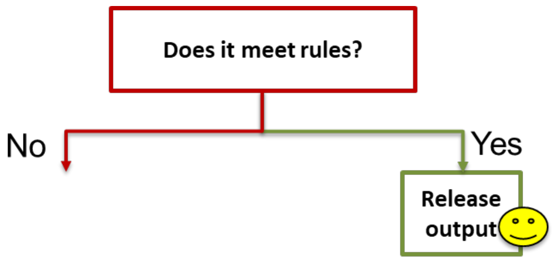


Figure 8 The DToD: meeting rules

If the rules are met, the output should be released. If the rules are not met, then the next tage depends on the output checking regime. If rules-based (for example, in an automatic system with no exceptions), then the output is rejected. If however this is a principles-based system and an exception is begin requested, the output checker needs to being evaluating the output.

9.3 Is it important?

The next stage is to determine whether the output checker should be spending time on this:

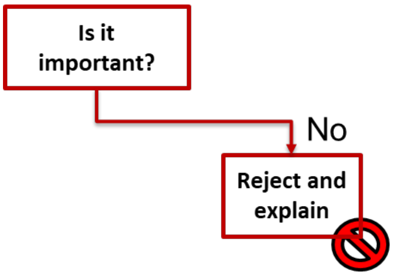


Figure 9 The DToD: ascertaining importance

The researcher should be able to give a reason why this exception is important (for example, requested by a journal). The output checker can assess that reason and judge whether it is a valid reason for looking at this in detail. If not ("I didn't have time to tidy up my outputs"), reject.

It is important that this question is asked before the next one, about disclosure. Output checkers have limited resources. The reason why the importance question is asked first is to determine whether the output checker should be spending time on assessing disclosure risk when other tasks are likely to be waiting.

9.4 Is it potentially disclosive?

Once the importance of the output has been agreed, the assessment of disclosure risk starts. Ideally:

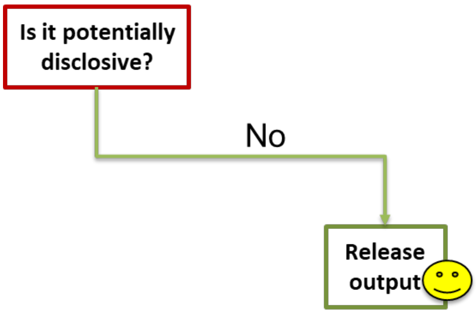


Figure 10 The DToD: Assessing disclosure risk

If there is no disclosure risk, release.

9.5 Are there special factors needing a policy decision?

There may be cases where unusual factors may indicate that there is a non-trivial disclosure risk. For example:

- a class disclosure leads to a significant policy outcome which should be published ("none of the children on Free School Meals had access to their own computer")
- the introduction of a new classification for medical cases causes differencing problems with outputs using the old classification; should the researcher be required to use the old classification, or should the new classification be used and the differencing risk accepted as part of the development of knowledge?

If there are none of these issues, the output can be rejected. However, there should be a clear line of seniority allowing the output checker to raise complex or unclear problems with others able to take a policy decision:

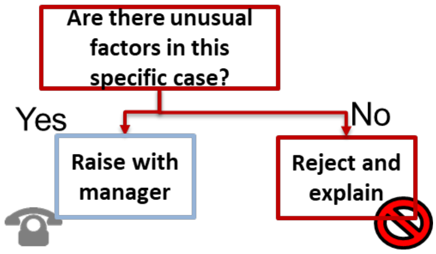


Figure 11 The DToD: unusual circumstances

In training, output checkers have generally found this helpful in structuring their analysis. The DToD can also usefully be shared with researchers to illustrate the thought processes of output checkers, and the importance of making sure that outputs are clearly described and intelligible.

10. Frequently asked questions about output checking

10.1 FAQs for researchers

To be done… have services got these already?

10.2 FAQs for output checkers

These FAQS are there to help output checkers deal with researcher queries. Researchers may also find the answer they are looking for here. This section is taken from Ritchie and Welpton (2015), which has a more detailed discussion for each item.

10.2.1 Researcher queries on disclosure risk

Yes, there's a lot of output, but it's all safe…

- ⇨ It is the researcher's responsibility to justify every release that they request. If we checked everything that a researcher produced, then we wouldn't be able to operate the service. Everybody needs to work efficiently. In addition, we shouldn't be releasing outputs if you don't require them because this potentially increases the risk of secondary disclosure.

This isn't disclosive, why can't I have it?

- ⇨ It is the researcher's responsibility to justify every release that they request. If we checked everything that a researcher produced, then we wouldn't be able to operate the service. Everybody needs to work efficiently. In addition, we shouldn't be releasing outputs if you don't require them because this potentially increases the risk of secondary disclosure.

Why is the amount of output important?

- ⇨ We don't have infinite resources. The more we check the results of person A, the longer person B has to wait for his or her results. Why should B be disadvantaged by A's lack of consideration? Are you happy to wait while we sort through someone else's messy output? We make sure that everyone keeps their output down so that we can give a better service to all, rather than the good researchers being penalised by the lazy ones.
- ⇨ Help us by providing only the output you require to have released. Everybody will receive their outputs back faster; we can continue to provide access as we can prove we are protecting the confidentiality of the data.

Why do results have to be checked?

- ⇨ If you've made a mistake in your output and disclose some confidential fact, you're in deep trouble – and so are we, as we gave you access. So, we both lose. I agree that your output is pretty likely to be safe, but checking it gives the extra ring of confidence – and we do know cases where it has helped us avoid producing results we shouldn't have. This process has allowed us to provide the service for so many years now and reassures data owners who are generally risk-averse about providing access to their data in the first place.

Will low cell counts actually lead to disclosure [of personal information]?

- ⇨ Probably not (although you need to be aware that certain things with extreme distributions such as rare illness or company data might be visible), but we don't have the time or resource to check every option; and it is not possible to prove non-disclosure in any case. However, assuming that low cell counts might be disclosive encourages us to build in a margin of error, so that we block many simple tabs and only check in detail when it's important. Just don't assume that every low cell count is also a breach.

10.2.2 Researcher queries on operational matters

When will I get my results back?

- ⇨ As soon as they are checked – which means, the better they are the faster they are likely to be returned. We are allowed to prioritise to do easy checks first, that is to say, results that are clearly explained and straightforward to assess. Messy results take more effort.

What makes you the expert on my output?

- ⇨ I'm not claiming to be; my concern is to understand what is being realised. We're clearly talking at loggerheads here; perhaps you can explain again? Can you provide an easy-to-understand methodology? This will help me to understand how you have produced your results. If I don't understand, I can't release it, because I'm taking on the risk.

I don't agree with your decision – who do I complain to?

- ⇨ (Response assumes you've already got some feedback from peers, and also that this is something other than just disagreeing about risk)
- ⇨ A: If you think you've been unfairly treated, I'm happy for you to talk to my supervisor.
- ⇨ B: If you think there's some evidence, I've not taken account of, I'm happy to pass your query on to my peers inside/outside the organisation

My client is pressurising me to get results out quicker – how do I respond?

- ⇨ It's important that you raised this. If they won't listen to you, we're happy to intercede and explain the circumstances under which access to data is granted. You should inform us immediately if you are put under such pressure. We have solutions in place to enable the supervisor to see results; but undue pressure may encourage risky behaviour that can lead to catastrophic consequences.

10.2.3 Other operational considerations

I trust this researcher – should I be more lenient with them?

- ⇨ No. If they are producing good output, you don't need to be lenient. If they are not producing good output, you need to educate them, and then not be lenient with them. Treating researchers differently stores up problems for yourself – what if you get challenged on it?

I don't trust this researcher – what do I do?

- ⇨ If they are producing bad output, educate them. If you don't trust them to produce genuine output, or use the facility properly, report to the RDC Manager. SDC isn't designed to cover bad faith research.

What should I do if the researchers are wasting my time with frivolous requests?

- ⇨ Educate, penalise by lowering priority – if nothing works, report to the facility manager for further action. PBOSDC can't run if people don't play by the operating rules (ironically)

What if the researcher refuses to accept my decision?

⇨ If they are unhappy about the process, refer upwards. If they are unhappy about the use of evidence, refer sideways. If they are unhappy about your judgement, turn it into a question of evidence and refer sideways again. Remain open to the possibility of error, and make sure the researcher knows that you are considering his/her complaint in an uncertain world. This gives you room to manoeuvre if it turns out you have made a poor call.

I don't understand what I'm being shown – what do I do?

⇨ Ask the researcher or ask a peer/supervisor. There is no shame in not knowing, but there is enormous reputational risk from pretending to know something you don't.

How much should I rely on what has been released before?

⇨ Unless previous releases are acknowledged as precedents, or used as examples of training guides, very little.

Should I check what has been released before?

⇨ In general no but do if you suspect there are multiple releases of near-identical outputs.

Can I come to an arrangement with the researchers to clear their output more effectively?

⇨ This is encouraged, as it demonstrates your awareness of researcher needs. For example, one researcher needed to produce a lot of tabular output. We agreed that we would (a) double the threshold limit from 10 to 20 units, but then (2) validate the program rather than all the individual outputs. This worked for both of us. Be careful though about setting precedents for operations.

An alternative scenario occurred where a researcher required a large tabular output, but to reduce the risk of secondary disclosure, he was provided with 'two chances' to get his output right: no more would be released.

I'm not comfortable checking some types of output –what should I do?

⇨ In the short term, talk to the researcher; if that still doesn't help, bring in an expert colleague. In the long term, learn about these outputs. What you shouldn't do is refuse them because you don't understand them (refusing them because the researcher refuses to explain is okay!).

In general, you should raise these issues with the RDC Manager. The chances are another output checker doesn't understand these outputs either. Then, the RDC Manager is aware and can provide appropriate training.

I don't understand the data – what should I do?

⇨ If we are talking about unsafe statistics, this is more important. Take advice from both the researcher (emphasise how they help you) and colleagues. You must be comfortable with the contents before release.

Do I have to give a reason why I refused something?

- Yes absolutely, and for two reasons. First, it is simply polite and good customer service to provide a reason. Imagine you visited a restaurant and asked for the steak that others are having, only to have the waiter tell you that you can't have it and walking away.

Secondly, by providing a reason, you have given the researcher the opportunity to amend their output, and also to learn from the experience. Education is an on-going process, and having made a mistake the first time, and understanding what the mistake is, they hopefully will not make the same mistake a second time.

I've released something I shouldn't have – what do I do?

- We all make mistakes. The best outcome is that you and everybody else understand this and learns from the mistake. In the worst case, it may become an 'information security event' but we strive to 'continuously improve.' In most cases, you should contact the researcher, explaining what the problem is, and consider how to prevent further release – which may not be possible. You should accept responsibility but may also need to point out to the researcher that inappropriate releases risk everyone's access, so it's in their interest to help address the problem.

I've refused something I shouldn't have – what do I do?

- Release it and apologise. Say that you've reconsidered the output and in light of this, you are happy to release it. You mustn't think anybody will think poorly of you for this: researchers and staff should work together and be sympathetic that we don't all get everything right all the time. Treat it as a learning experience, as you would expect of them if they submit an output that really can't be released.

What if my staff member makes a mistake?

- As a facility manager, you are responsible for your team undertaking output checking. The first thing is to assess the extent of the mistake: has this in your view, led to an 'information security event'? If so, you'll need to take appropriate action, in the form of a 'Corrective Action Plan' (which details what happened, why it happened, what you can implement to prevent it from happening again and with a time frame).

  Whether or not this is a disclosive release, you should consider the following:

  - Is this a one-off mistake by one person (then educate)
  - Has this occurred more than once by the same person (then it's a potential line management issue)
  - Has this occurred by more than one person (then something systematic is going on which needs to be resolved as a team)

What if my colleague makes a mistake?

- Hopefully, this can be discussed between yourself and the colleague, but if you are uncomfortable, then you should speak to the RDC Manager. They should undertake a monthly 'quality assurance': reviewing a selection of outputs released in the previous month. They may decide to focus on the outputs released by a particular member of staff if concerns have been raised.

What if the researcher presents a case for release, I don't understand?

- This is simple: just ask them. You'll need to understand:
  - What the results show: can the researcher provide you with a sentence interpreting the statistics, e.g., 'this shows that x and a positive effect on y.'

    - How the results were calculated: the methodology used.
- Normally, researchers are very pleased to explain this to you. Researchers who aren't keen should be viewed suspiciously: why aren't they happy to explain, are they hiding something?

How should I prioritise when I have multiple outputs to check?

- ⇨ Option 1 is first-come-first served, which is fair and equitable, but doesn't reward those who produce good output. We would suggest you prioritise as follows:
    - (1) good and quick
    - (2) good and slow
    - (3) poor and quick
    - (4) poor and slow
    - (5) chancers and idiots.
- ⇨ Make sure you explain the order to researchers, otherwise you don't see the benefit.

11. Other guides and manuals

These are guidelines already out there which provide specific advice for output checking (both output checkers and researchers).

11.1 Introductory guidance for OSDC

These are introductory pieces on SDC, suitable for someone with no knowledge and needing a short introduction.

Title	Ensuring the confidentiality of statistical outputs from the ADRN
Authors	Lowthian, Phillip; Ritchie, Felix
Date	2017
Description	Narrative introduction to very basic OSDC
Coverage	• What is SDC? • What is output based SDC? • Actual vs potential disclosure • Principles based OSDC • How much risk is there in research outputs? • Making PBOSDC work in the ADRN • Working with researchers • Researcher training • Ensuring confidentiality and flexibility
Citation	Lowthian, P., Ritchie, F., Mackay, E., & Elliot, M. (2017). Ensuring the confidentiality of statistical outputs from the ADRN
Link	Ensuring the confidentiality of statistical outputs from the ADRN

Title	Safe Researcher Training slide deck
Authors	Green, Elizabeth; Ritchie, Felix; Office for National Statistics; SRT Expert Group
Date	2019 (web version)
Description	Slide deck for training researchers, with extensive notes
Coverage	• What is SDC? • Small counts • Class disclosure • Structural zeros • Cell suppression • Rounding • Output redesign • Dominance • Ranks, Maxima, Minima • Differencing • SDC and statistical quality • High vs Low review statistics
Citation	ONS (2019). Safe Researcher Training (v0.13 updated 2022). September.
Link	saferesearchertraining.org/SRT_slides.html

Title	Eurostat: Statistical Disclosure Control (Memobust Handbook)
Authors	Willenborg, L; de Wolf, P-P; Eurostat
Date	2014
Description	Short overview on OSDC, designed for producers of official statistics
Coverage	• Tables vs microdata.

	• Tabular data. • Probability of disclosure versus information loss. • User needs and SDC. • Data access. • Design issues. • Software tools. • Decision tree of methods [not complete?].
Citation	Willenborg, L. and de Wolf, P.P. (2014) *Statistical Disclosure Control – Main Module*. Netherlands: Memobust.
Link	https://ec.europa.eu/eurostat/cros/system/files/Statistical Disclosure Control-01-T-Main Module v1.0.pdf

11.2 General guides to OSDC for researchers

The guides listed here have been developed to support researchers using confidential microdata. Note that secure facility staff in the SACRO network identified the SDAP manual as the most helpful, hence why it is listed first. Others are in alphabetical order.

Title	SDAP Handbook on Statistical Disclosure Control for Outputs
Authors	Griffiths, Emily; Greci, Carlotta; Kortrotsios, Yannis; Parker, Simon; Scott, James; Welpton, Richard; Wolters, Arne; Woods, Christine.
Date	2019
Description	General-purpose guide for researchers and output checkers covering a rang of outputs and practical guidance
Coverage	• Statistical risk • What is SDC? • Risk assessment • Statistical risk: Principles and rules • Introduction to SDC • Descriptive statistics • Percentiles • Histograms • Box plots • Correlation coefficients • Factor analysis • Indices • Scatter plots • Symmetry plots • Decision trees and exclusion criteria • Survival analysis; Kaplan-Meier curve • Spatial analysis (maps) • Gini coefficients • Concentration ratios • Regressions • Residuals • Margin plots • Test statistics • Implementing SDC as an organisation • Managing analysts • Managing expectations

Citation	Welpton, Richard (2019). SDC Handbook. figshare. Book. https://doi.org/10.6084/m9.figshare.9958520.v1 [check]
Link	SDC Handbook – Secure Data Access Professionals (SDAP) (securedatagroup.org)

Title	ABS Data Confidentiality Guide
Authors	Australian Bureau of Statistics
Date	2021
Description	General purpose guide for researchers
Coverage	• Safely releasing valuable data. • Confidentiality (Obligation to maintain confidentiality). • Legal obligations. • Contextual approaches. • Re-identification. • Administrative data. • Integrated data sets. • Big Data analytics. • Reidentification in aggregate data and microdata. • Five Safes. • Tables and disclosure risks: Frequency tables, Magnitude tables. • How to identify at risk cells. • Frequency rule. • Cell dominance rule. • P% rule. • Tabular data. • Suppression. • Data modification. • Hierarchical data. • Treating microdata. • Assessing disclosure risks. • ABS microdata
Citation	Australian Bureau of Statistics. (2021, November 8). *Use of ABS microdata and impact on research quality*. ABS.
Link	https://www.abs.gov.au/about/data-services/data-confidentiality-guide

Title	CENEX SDC Handbook
Authors	Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Naylor, Giovanni Seri, Peter-Paul De Wolf
Date	2010
Description	Eurostat-commissioned guide to all aspects of input and output SDC, intended to reflect the then state of knowledge. Incorporates Brandt et al (2010) on outputs. Note the team produced a similar version book form (Wiley).
Coverage	• Regulations: (Ethical codes, Laws). • Microdata: (Roadmap to microdata release, Risk assessment, Microdata protection methods, Information loss in microdata protection, Software). • Magnitude tabular data: (Disclosure control concepts, the τ-ARGUS implementation of cell-suppression, concepts of secondary cell suppression algorithms in τ-ARGUS, Controlled tabular adjustments).

	• Frequency tables: (Disclosure risks, Methods, Rounding, Information loss, Software). • Remote access.
Citation	Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Nayor, J., Nordholt, E.S., Seri, G., De Wolf, P.P. (2010) *Handbook on Statistical Disclosure Control. V 1.2*. Netherlands: EuroStat.
Link	https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

Title	DwB: Guidelines for the checking of output based on microdata research
Authors	Steven Bond, Maurice Brandt, and Peter-Paul de Wolf
Date	2012
Description	General guide, very closely based on Brandt et al (2010)
Coverage	• Principles-based model. • Rule-of-thumb model. • Descriptive statistics (Frequency table, Magnitude tables, Maxima, Minima, Percentiles, Modes, Means, Indices, Rations, Indicators, Concentration ratios, higher moments of distributions, graphs). • Correlation and regression analysis (linear, non-linear, estimation, residuals, summary and test statistics, correspondence analysis). • Organisational/procedural aspects. • Legal basis. • Access requests. • Responsibility for quality. • Guide for number/speed of checkers. Researcher training
Citation	Bond, S., Brandt, M. and de Wolf, P.P. (2012) *Guidelines for the checking of output based on microdata research*. Luxembourg: Eurostat & DwB
Link	https://ec.europa.eu/Eurostat/cros/content/recommendations-protection-census-data_en

Title	Eurostat: How to use microdata properly
Authors	Ritchie, Felix; Eurostat
Date	2021
Description	Self-study material about OSDC and handling confidential microdata
Coverage	• Perceptions about research use of data. • Restrictions on data access: (intruder model and human model). • Ways of providing data access. • A framework for data access. • When things go wrong. • Five safes. • Protection of tables. • Class disclosure. • Protection of graphs. • Safe and Unsafe statistics. • Dealing with unsafe statistics. • Software support.
Citation	EuroStat (2021) *How to use microdata properly*. Luxembourg: EuroStat.
Link	https://biblioguias.cepal.org/eurostat/studymaterial

Title	ONS: Researcher output clearance guidance
Authors	Flavell J.; Lock A.; Greenwood C.; Office for National Statistics (UK)
Date	2022
Description	Guide for researchers using the Secure Research Service, but focusing more on process than outputs
Coverage	• What is an output? • How is an output cleared? • Levels of clearance • What doesn't need to be cleared • Incidents and breaches • Process Diagrams
Citation	ONS (2019) SRS Researcher output clearance guidance. Office for National Statistics.
Link	https://www.ons.gov.uk/file?uri=/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/gettingyourresearchoutputsapproved/researcheroutputguidancev2.6.pdf

Title	ONS: SRS Output checking guidance document
Authors	Office for National Statistics (UK)
Date	2022
Description	Guide for output checkers in the Secure Research Service
Coverage	• Clearance types • General output guidance • 'Safe ' and 'Unsafe' outputs • Default SDC 'rules of thumb' • File types • Frequency tables • Other tables • Low counts • Zeros • Suppression • Rounding • Reformatting • Class disclosure • Structural zeros • Secondary disclosure • Dominance • Statistics • Mean • Percentages • Weighted counts • Mode, minimum and maximum • Medians, quartiles, deciles and percentiles • Ratios • Odds ratios • Graphs (Line, bar, scatter, histograms, boxplot, violin plot) • Regressions and modelling • Coefficients, Margin plots and test statistics

	• Residuals • Maps and spatial analysis • Geographics • Code files
Citation	ONS (2019) SRS Output checking guidance. Office for National Statistics.
Link	https://www.ons.gov.uk/file?uri=/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice/gettingyourresearchoutputsapproved/srsoutputcheckingguidance.pdf

11.3 OSDC guidance produced for official statistics

These guides were produced to support the production of official statistics. As such, they contain relevant information and a different perspective; however, researchers should be aware that these are designed for large organisations with formal approval processes, using correction methods (eg secondary suppression) which are no recommended for researchers.

Title	GSS: Guidance for tables produced from administrative sources
Authors	Government Statistical Service (UK)
Date	2014
Description	Introduction to tabular data protection from administrative data
Coverage	• Key steps. • Guidance on administrative tables. • Implementation and evaluation of tabular data from administrative services. • Responsibilities. • Determining user requirements. • Understanding key characteristics of data and required outputs. • Circumstances where disclosure is likely and how to manage this. • Disclosure risk and breach of statistical obligations. • Selecting SDC rules and methods.
Citation	GSS (2014) GSS/GSR *Disclosure Control Guidance for Tables Produced from Administrative Sources.* UK: Government Statistical Service
Link	https://gss.civilservice.gov.uk/wp-content/uploads/2018/03/Guidance-for-tables-produced-from-administrative-sources-4.pdf

Title	GSS: Guidance for tables produced from surveys
Authors	Government Statistical Service (UK)
Date	2014
Description	Introduction to tabular data protection
Coverage	• Key steps • Guidance on tabular data from surveys • Implementation and Evaluation • Responsibilities • Relevant legislation • GSS policy • Statements made to respondents. • Trust of respondents

	• Guidance for tables produced from social surveys. • Determining user requirements • Understanding key characteristics of data, and required outputs • Circumstances where disclosure is likely and how to manage this • Disclosure risk and breach of statistical obligations • Selecting SDC rules and methods • Implementation issues and concerns
Citation	GSS (2014) *GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys*. UK: Government Statistical Service.
Link	https://gss.civilservice.gov.uk/wp-content/uploads/2018/03/Guidance-for-tables-produced-from-surveys-4.pdf

11.4 General guides to process

These are general comments on the process of setting up and managing OSDC processes.

Title	Eurostat: How to be a safe researcher
Authors	Eurostat
Date	No date
Description	Guide from Eurostat on managing access to research data
Coverage	• About Eurostat and microdata access. • ECLAC Status. • ECLAC roles and responsibilities. • Who can and can't access data. • Confidentiality agreement. • Breach of confidentiality. • Requesting access to microdata. • Using and managing microdata. • Eurostat datasets.
Citation	Eurostat (n.d.) *How to be a safe researcher*. Luxembourg: Eurostat.
Link	https://biblioguias.cepal.org/Eurostat/documents

Title	OCHA: Statistical Disclosure Control Guidance
Authors	United Nations Office for the Coordination of Humanitarian Affairs (OCHA)
Date	2019
Description	Sohrt discussion of input versus output SDC and its role in project design
Coverage	• Humanitarian microdata • Re-identification and disclosure risk: (Identity disclosure, Attribute disclosure, Inferential disclosure) • Statistical Disclosure Control: (Risk assessment, Application of SDC methods, Reassessing risk and quantifying information loss) • Application of SDC in humanitarian data management
Citation	OCHA. (2019). Statistical Disclosure Control Guidance. Retrieved from https://centre.humdata.org/wp-content/uploads/2019/07/guidance_note_sdc.pdf
Link	https://centre.humdata.org/wp-content/uploads/2019/07/guidance_note_sdc.pdf

12. References

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

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Maurizio, L., Mol, J., Ritchie, F., Seri, G., & Welpton, R. (2010). *Guidelines for the checking of output based on microdata research. Final Report of ESSnet Sub-Group on Output SDC*. Eurostat. https://uwe-repository.worktribe.com/output/983615/

Derrick, B., Green, E., Ritchie, F., & White, P. (2022a). The Risk of Disclosure When Reporting Commonly Used Univariate Statistics. *Privacy in Statistical Databases 2022*, *13463 LNCS*, 119–129. https://doi.org/10.1007/978-3-031-13945-1_9

Derrick, B., Green, E., Ritchie, F., & White, P. (2022b). Disclosure risks in odds ratios and logistic regression. *Scottish Economic Society Annual Conference 2022.* https://uwe-repository.worktribe.com/output/9853160

Derrick, B., Green, E., Ritchie, F., & White, P. (2023). Towards a comprehensive theory and practice of output SDC. *UNECE/Eurostat Expert Group on Statistical Data Confidentiality, 2023.* https://unece.org/sites/default/files/2023-08/SDC2023_S5_2_UWE_Ritchie_D.pdf

Desai, T., & Ritchie, F. (2009, December). Effective researcher management. *UNECE Worksession on Statistical Data Confidentiality 2009*. https://uwe-repository.worktribe.com/output/989767/

Green, E., & Ritchie, F. (2021). Statistical disclosure control for HESA: Part 1: Review of SDC theory. Higher Education Statistics Agency (HESA) https://uwe-repository.worktribe.com/output/10621148/

Green, E., Ritchie, F., Bradley, P., & Parry, G. (2021). Financial resilience, income dependence and organisational survival in UK charities. *Voluntas*, *32*(5), 992–1008. https://doi.org/10.1007/S11266-020-00311-9

Green, E., Ritchie, F., Tava, F., Ashford, W., & Ferrer Breda, P. (2021). *The present and future of confidential microdata access: Post-workshop report.* University of the West of England Bristol. https://uwe-repository.worktribe.com/output/8175728/

Griffiths, E., Greci, C., Kortrotsios, Y., Parker, S., Scott, J., Welpton, R., Wolters, A., & Woods, C. (2019). Handbook on Statistical Disclosure Control for Outputs. *Safe Data Access Professionals Working Group*. https://doi.org/10.6084/m9.figshare.9958520

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Schulte Nordholt, E., Seri, G., & De Wolf, P. (2010). Handbook on Statistical Disclosure Control. *ESSNet SDC*. https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf

ONS (2019) *Safe Researcher Training; canonical slides* http://www.saferesearchertraining.org/SRT_slides.html

Ritchie, F. (2006). Disclosure control of analytical outputs. Mimeo, ONS. Republished as WISERD working Paper no 5, 2011. https://uwe-repository.worktribe.com/output/957303

Ritchie F. (2007). Disclosure detection in research environments in practice. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*. https://uwe-repository.worktribe.com/output/1023059

Ritchie, F. (2014). Operationalising "safe statistics": the case of linear regression. *UWE Working Papers in Economics No 14/10*. https://uwe-repository.worktribe.com/output/808629/

Ritchie, F. (2017). *The "Five Safes": A framework for planning, designing and evaluating data access solutions*. Paper presented at Data for Policy 2017, London, UK. September http://dx.doi.org/10.5281/zenodo.897821

Ritchie F. (2019). Analyzing the disclosure risk of regression coefficients. *Transactions on Data Privacy*, *12*, 145–173. http://www.tdp.cat/issues16/tdp.a303a18.pdf

Ritchie F. and Welpton R. (2015) *Operationalising principles-based output statistical disclosure control*. Working paper, University of the West of England Bristol. https://uwe-repository.worktribe.com/output/8073309

Smith, J. E., Clark, A. R., Staggemeier, A. T. and Serpell, M. C. (2012). A Genetic Approach to Statistical Disclosure Control. IEEE Transactions on Evolutionary Computation, 16(3), 431-441. http://dx.doi.org/10.1109/TEVC.2011.2159271

Smith, J., Preen, R., Albashir, M., Ritchie, F., Green, E., Davy, S., Stokes, P., & Bacon, S. (2023). SACRO: Semi-Automated Checking Of Research Outputs. *UNECE/Eurostat Expert Group on Statistical Data Confidentiality, 2023.* https://uwe-repository.worktribe.com/output/11060964/

Part IV Technical appendices

Appendix 1 Note on survival tables

12.1 Summary

The current guidance for survival tables and Kaplan-Meier (KM) graphs is that researchers should demonstrate that each step in the table contains at least the threshold number of observations.

This paper argues that the actual disclosure risk in the table and or graph is negligible. While retaining the classification as 'unsafe', because a notional differencing risk exists, within the unsafe outputs it is classified as 'very low risk'. Both manual and automatic output checking guidelines should focus on the total and residual number of observations, and on outliers.

The summary table to be included in the DRAGoN manual is:

Examples of type	Survival tables, Kaplan-Meier graphs
Safe or unsafe?	Unsafe
Risk level	Very low
Risk elements	Class disclosure Extreme survival cases
Checks to be made	Initial and final thresholds, extreme survival values
Appropriate responses	Suppression (via right-censoring)
Covered in automatic tools	SACRO [tbd]
Modelling	Limited
Key text(s)	[This paper]

12.2 Tabular outputs

Survival tables (where 'survival' does not just mean life outcomes, but other events that persist and then stop) are different from other frequencies tables because there is an implicit relationship between cells.

Date	Survivors		Date	Deaths
01.03.20	1,369		01.03.20	
08.03.30	1,298	Implicit death rate -->	08.03.30	71
15.03.20	1,281		15.03.20	17
22.03.20	1,225		22.03.20	56
29.03 20	1,216		29.03 20	9
05.04.20	1,199		05.04.20	17
…			…	

Table 29 Implied columns in survival data

The same information could be expressed as survival days, or as rates from an original population, or rates from survivors to date. Deaths could also be expressed as a proportion of the original population

Day	Surviving	Deaths	Survival rate	Death rate	Hazard rate
0	2300	(original population)			
1	2286	14	99%	1%	1%
2	2131	155	93%	7%	7%
3	1930	201	84%	16%	9%

4	1565	365	68%	32%	19%
5	1532	33	67%	33%	2%
6	1322	210	57%	43%	14%
7	1287	35	56%	44%	3%
8	1255	32	55%	45%	2%
9	1023	232	44%	56%	18%
10	854	169	37%	63%	17%
11	834	20	36%	64%	2%
12	690	144	30%	70%	17%
13	591	99	26%	74%	14%
14	564	27	25%	75%	5%
15	512	52	22%	78%	9%

Table 30 Alternative restatements of the survival table

If the original population is known, any of these columns can be created from any of the others (assuming there are sufficient decimal points on the rates).

For survival tables the row categories (in tables 1 and 2) are derived from the data rather than having any external validity (eg compared to the usual table categories, such education levels). This means that the row categories themselves are uninformative (with the exception of outliers – see below). Consider the table above. It states that there were 2300 in the original population. These individuals have the same characteristics – that is why they are in the same table. If the variable of interest was survival time, the table indicates that 20 people exited the study on day 11, but there is nothing to distinguish these individuals from those who exited on other days.

As an alternative, consider what a single exit represents in, say, a table showing survival rates of males aged 55-59. The table confirms that there was an exit on that date was for a male aged 55-59; but these also exist for every other non-zero date. It is of course possible to confirm identification: if an individual was known to have exited on day X, and there is only one exit on that day, that it might reasonably be assumed to relate to that individual – if the individual is known to be in the dataset (which might be the case in population level datasets).[4]

However, in the case of outliers included in the tables, the categories are likely to be informative and only holding a single value. Suppose the above table continues…

Day	Surviving	Deaths	Survival rate	Death rate	Hazard rate
0	2300	(original population)			
:	:	:	:	:	:
:	:	:	:	:	:
87	4	1	0%	100%	20%
92	3	1	0%	100%	25%
94	2	1	0%	100%	33%
112	1	1	0%	100%	50%

[4] It depends what we're classing as a risk here - is it (as is commonly used in training) that ANYONE identifying ANYTHING is a problem (so if I can pick out that I myself am in the dataset, this is an issue)? Or that is anyone ELSE could pick me out AND learn something about me, this is what we're concerned with? We generally assume the latter - the former essentially leads down a rabbit hole of nothing every being published. However, identifying that you are the one person being treated on a day does provide a form of class disclosure about others..

206	0	1	0%	100%	100%

Table 31 Extreme values

In this case, knowing that an individual stayed in the study for much longer than others but not knowing exactly how long, this could potentially reveal information about the exact length of stay.

Differencing is a theoretical concern. Imagine a researcher producing multiple tabulations by gender, age, ethnicity, work status, treatment and pre-diagnosed condition, plus a tabulation of the entire dataset. If there were a date on which just one individual exited the study, then it would be possible to identify the characteristics of the individual who exited on that date (eg "male, age 55-59, not white, in employment, liver failure, no pre-conditions"). Note that this requires the survival table for the study population also to be published (or failing that, all possible combinations to be shown), and for that to show just one individual at that exit point. Otherwise, it is not possible to identify a single-exit in one tabulation with a single-exit in another.

Class disclosure is a potential consideration: "all the patients recovered within 60 days". As is usual with class disclosure, this is entirely context-dependent. If there are no survivors at period end, then membership of the study group is potentially informative that the respondent exited the study, although there is no clue when other than the last survival time.

Right-censoring (ie having some individuals still in the study) cannot increase disclosure risk. It is more likely to reduce it as it adds uncertainty about exact survival times and removes extreme outliers.

12.3 Risk assessment

Survival tables should be classified as 'unsafe' because there are the three problems of low numbers, class disclosure and outliers. However, survival tables are a very low risk output for several reasons

- Only one piece of information is presented (the rows) and this determines the category label
- All individuals on the graph share the same characteristic apart from exit date
- The outcome is the identifying variable: if respondent X passes away on 10th June 2023 and there is only one death on that day, this identifies respondent X; but there is no independent information which allows one to say "this is respondent X and I have identified that he/she passed away on 10th June" (in other words, identification risk but no attribution risk)
- Moreover, the time units are likely to be relative to first identification (eg number of days since infection); hence respondents are likely to have different starting times (so 'day 6' will mean different absolute dates for different patients)
- The starting population is likely to be large, even if the number dropping out at each period is small; otherwise the steps are of very limited statistical value
- If one knew an individual's exit date, and that they were in the dataset, and they were the only exit on that date, then one could associate the characteristics of that group with the individual; but this is quite a high bar

This also assumes 100% confidence in the accuracy of the exit date. This is unlikely to be warranted for frequent exit dates (eg days) where administrative errors are likely to come into play. Less frequent exit dates may be more accurate but may also have more exits. If exit dates are relative to recruitment date rather than absolute then identifying an individual requires 100% confidence in identifying the recruitment date as well the exit date[5].

There are two meaningful risks: extreme outliers and class disclosure.

Extreme outliers should be easily identifiable. Moreover, even if left in, they are less likely to be very informative than outliers in other statistics, as there is no independent information in the table which could be used or extracted to find out about the respondent. Again, the fact that everyone in the table has the same characteristics limits the risk. For an outlier to be informative, an intruder would have to be confident that the outlier is the person of interest (by knowing for example, a rough value to exit date). Right-censoring removes the extreme outliers.

Class disclosure could occur either because no-one exits before a certain date, or exits after a certain date, and that date is informative. An 'informative' date would need to be one the last exit date is unexpectedly early. Again, if there is right-censoring there is no class disclosure (ie just as for other statistics where the range of values is trimmed).

To illustrate the low risk, consider what factors, other than outliers and class disclosure, would make a survival table problematic for subject X:

- Exit date (relative or absolute) for X is known with certainty
- Exit date is unique to X
- X has been confirmed as part of the tabulated population

In this case the characteristics of the table can be ascribed to X. However, this is a high bar, even with a known population. Moreover, it is far more likely that one would know the characteristics of X with certainty and not the exit date, rather than the other way round.

In short, while survival tables should be considered 'unsafe' in that they do need to be checked, they should be seen as very low risk: they contain almost no new information, and the main disclosure risk comes from class disclosure and outliers, which are easily identified.

It may be possible to define rules for what is an 'informative' end date or outlier. A well-populated table up to the end point/outlier is perhaps an indication that there are no disclosure risks. This requires some empirical study.

12.4 Checking survival tables

12.4.1 Manual checks

The manual output checker should confirm that

- Relative dates rather than absolute dates are used
- The initial population exceeds the usual threshold applied to frequency tables (note: this is for consistency with frequency tables, as this is effectively a cell count)
- If right-censored, the final population exceeds the usual threshold
- If not right-censored
    - the final exit date is not informative (is it unexpected/unreasonable?)

[5] All data are of course subject to error. The difference between survival tables and others is that the respondent does not supply the exit date, but that it is observed by the data administrator, providing a n additional opportunity for error beyond the usual factors of recall error or mistake.

- there are no extreme outliers which could be reasonably associated with individuals (is there a very substantial gap such that one might guess the individual being targeted is 'in that region'?)

The first two rules provide consistency with regular frequency table rules. In practice, we would generally expect the first to be passed easily; and the second is very uninformative.

12.5 Automatic checks

For an automatic checking program, the requirements for approval are

- The initial population exceeds the usual threshold applied to frequency tables
- The table is right-censored with the final population exceeds the usual threshold

If not right-censored, then this needs to be marked as 'fail' in a rules-based model, or 'for review' in a principles-based framework. Alternatively, the program could force right-censoring at the usual threshold.

12.6 Graphical outputs – the Kaplan-Meier curve

Survival analyses are typically shown as a Kaplan-Meier cure, which is a re-presentation of the tabular data. Consider the following, which is an extended version of the above table:

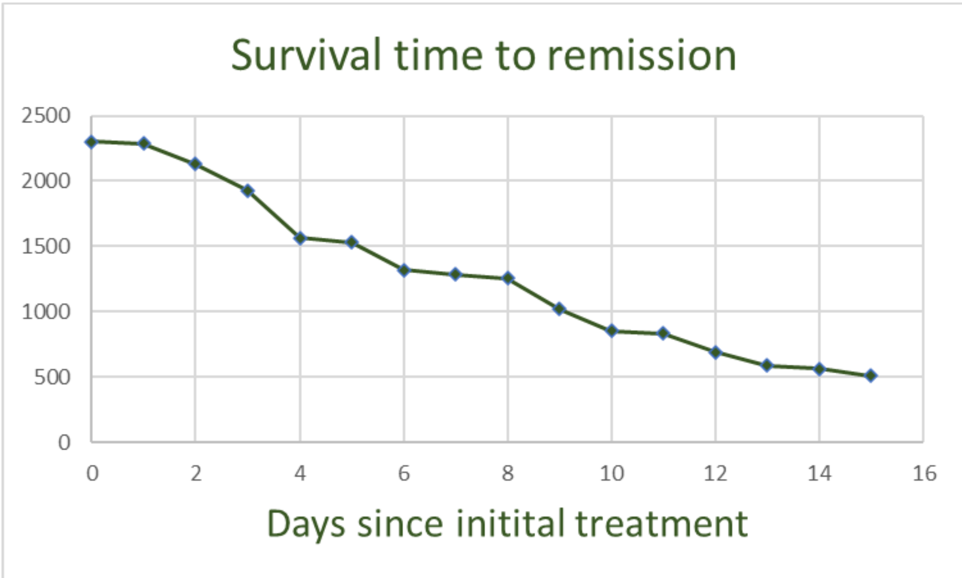


Figure 12 Kaplan-Meier example

In theory the gaps between survival points can be read off the graph. SDAP(2018) advises that the survival table should be requested for release as well. That may work in this case but consider the full graph:

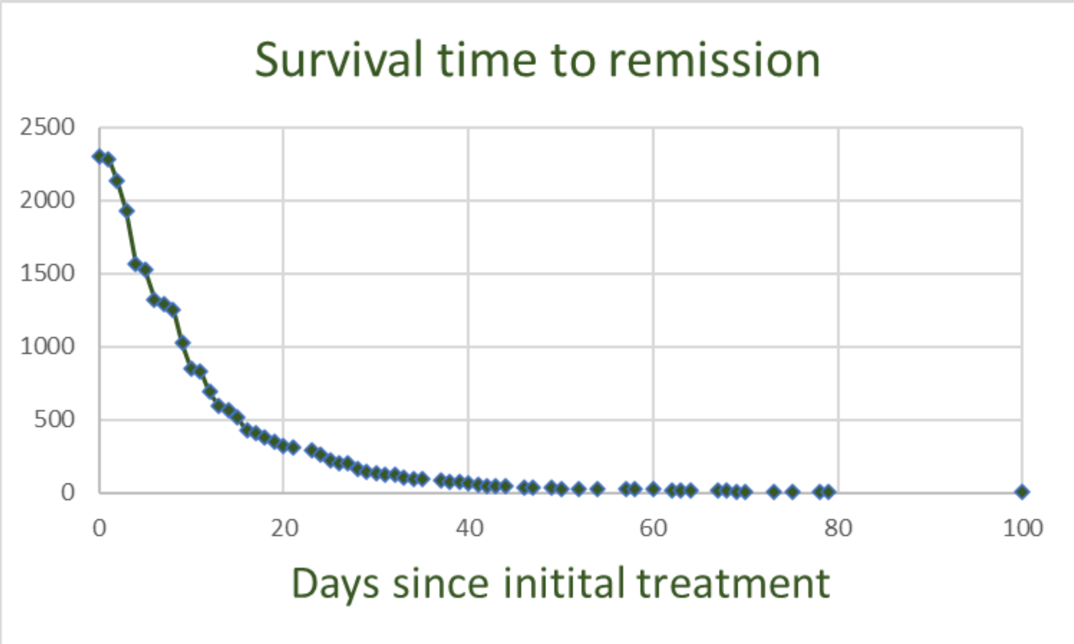


Figure 13 Kaplan-Meier example – full data

With 100 data points, this will take a long time to review. Unfortunately, a graph like Figure 2 is more likely to be requested than Figure 1. It's also worth considering what the release of the table means. The graph can be unpicked, yes; but with some effort. The table is easy to analyse for low numbers. So releasing the table as well as the graph increases practical risk, even if it does not increase theoretical risk.

However, if we start from the conjecture that the table is low risk to start with, and the graph is a less readable version of the table, we would recommend that in most cases it is not necessary to see the full survival table. Instead, ask the researcher to confirm

1. Starting observations
2. Final observations

KM graphs are more commonly presented as percentages, not numbers as shown above. This adds an additional element of protection. The disclosiveness of the graph is reduced, as the data points no longer just rely upon pixel resolution but also the number of decimal points used for the calculation.

12.7 Regression analyses

There are a number of regression models associated with hazard and survival data, including the Cox proportion and partial hazards models. The analysis of this has still to be done, but we expect to class them with other linear and non-linear estimates model as 'safe' with the same restrictions.

1) Check residual degrees of freedom exceed threshold
2) Check model is not fully saturated
3) Block estimation on a single dummy explanatory variable

The first two are very rare in general, designed to ensure this is a genuine regression, and can be ignored without compromising output security. Case (3) is a special case of (2), and the only one likely to occur in practice, and it is easy to check for manual and automatic checkers.

Appendix 2 Development notes on selected statistics

Statistical hypothesis tests

We note that the simpler tests can all be recast as regression models. Therefore for now, we apply the regression rules and impose a DoF restriction (where this is 'residual degrees of freedom' ie number of observations less restrictions). See however 'binary regressions' below.

Assessment

Manual and automatic:

- Check minimum degrees of freedom

Position: percentiles

Analysis

Position is reporting a value for the xth percentile of the distribution eg median is the 50th percentile. Effectively that means telling you something about the observations on either side of that value. Hence, to be consistent with magnitude tables one should have at least the minimum threshold on either side of the projected value:

The xth percentile is allowed if, for N observations and a threshold of T,

$$min(x, 1-x)*N>=T$$

For example:

N	Threshold	Allowed?		
		10%	50%	99%
10000	20	Yes	Yes	Yes
500	10	Yes	Yes	No
40	5	No	Yes	No

Table 32 Percentile rules

Where inter-percentile ranges are being used, we could require that the two percentile values are presented separately. This is purely to simplify checking – gap between two percentiles should be less disclosive than the values at a specific percentile. However, this leads into the problem that you could calculate individual values eg on 100 observations you quote the 10% and 11% boundaries ie one observation.

On the other hand, checking that the range meets the threshold is also not enough eg for 100 observations, the 1%-99% range would provide information about the gap between lowest and highest values.

So the formal rule should be:

For the set of percentiles x_1, $x_2..x_k$, N observations and a threshold of T, the x_ith percentile and x_i-x_j interquartile range is allowed if, for all i and j,

$$min(x_i, 1-x_i)*N>=T \text{ AND } (x_i-x_j)*N >=T$$

For example, continuing the above case and just checking the inter-percentile ranges:

N	Threshold	Allowed?		
		10%-90%	25%-75%	1%-99%
10000	20	(1000-9000) Yes	(2500-7500) Yes	(100-9900) Yes
500	10	(50-450) Yes	(125-375) Yes	(5-495) Yes
40	5	(4-36) Yes	(10-30) Yes	(1-39) Yes

Table 33 Inter-percentile range

As can be seen, it is quite difficult for the difference measure to fail the threshold.

Riskiness

While technically an **unsafe statistic**, it is low risk. The disclosure relies upon being able to identify where in the ranking an observation is – difficult to do without have the ranking already.

In addition, the reported percentile may not exist in the dataset. Some packages will calculate an intermediate value eg for the values

Position	1	2	3	4	5	6	7	8
Value	0	1	12	13	16	19	23	42

Table 34 Defining the median

The median value could either be presented as 13, 16, or 14.5 (13+16/2). This provides extra protection.

There is a class disclosure issue if the percentile value is common to the class eg if everyone in the bottom 10% earns the minimum wage. However, because ranking determines your inclusion in that class, we should be less concerned than we are with class disclosure normally ie the lowest-paid 10% earning the minimum wage is less informative than all school-leavers with no qualifications earnings the minimum wage. If the research paper showed that there was exact agreement between these two the issue would be that all the school leavers are in the bottom 10%, not that the bottom 10% earns the minimum wage (see example in the SRT on class disclosure).

Assessment

Manual:

- Check thresholds on both sides (n%, 1-n%)
- Check difference thresholds in inter-percentile ranges
- Simple check – is 1% x N > threshold? If so, everything else should be fine

Automatic:

- Check thresholds on both sides (n%, 1-n%); assume inter-percentiles will be counted from these and these will exceed threshold (OSDC assume people are not deliberately trying to cheat)

Extreme values: maxima and minima

Analysis

Technically, we could consider max and min as part of the 'position' spectrum, just dead end. They would always fail the threshold rule (n% or 1-n% < threshold) unless there were multiple people at the threshold, in which case it may not matter.

This seems like neatness for the sake of it, and it ignores that max/min provide a clear class disclosure. Again, *technically*, a position is a class disclosure (if you know someone is in the bottom half of the distribution, then the median tells you the max value); but you still need to know something about ranks. The point about max/min is that you only need to know someone is in the dataset for the disclosure to appear.

Max and min are also more likely to be structural, as they reflect the range of the data, not an intermediate point. So we continue to treat max/min as a separate class.

Riskiness

These are unsafe and relatively high risk – if disclosive. However, assessing for disclosure should be a straightforward assessment of whether the extreme value is of concern (either no meaningful, or structural, such as a max percentage being 100%). The problem for SACRO is that this is not easily automatable.

Assessment

Manual:

- Allow if not informative about individuals
- Factors to consider: potential for many individuals to have the same value; whether a genuine extreme value; whether genuinely structural

Automatic:

- Block and only allow as exception

The researcher should be able to answer the manual queries relatively easily.

Shape: standard deviation, skewness, kurtosis

Analysis

Shape statistics seem similar to correlations, in that they involve summed squared (cubed, fourth power) deviations which cannot be directly unpicked. However, because these are univariate, there is a potential differencing risk. Consider this, extracted from 'Little White Lies' (Derrick et al, 2022)

> if a new sample point x_{n+1} is introduced into a sample, then the new sample variance based on n+1 observations with n degrees of freedom, say s_{n+1}^2 , is given by
>
> $$ns_{n+1}^2 = (n-1)s_n^2 + (x_{n+1} - \bar{x}_{n+1})(x_{n+1} - \bar{x}_n)$$
>
> where s_n^2 is the sample variance for the original n observations based on $(n-1)$ degrees of freedom, $\bar{x}_n$ is the sample mean based on n observations and $\bar{x}_{n+1}$ is the sample mean based on $n+1$ observations.

From this it is possible to derive (using the fact that $\bar{x}_{n+1} = (n\bar{x}_{n+1} + x_{n+1})/(n+1)$) that

$$x_{n+1} = 1 \pm \sqrt{\bar{x}_n^2 - \frac{(n+1)}{n}(ns_{n+1}^2 - (n-1)s_n^2)}$$

Probably. Anyway, the point is that in theory you only need the two SDs and one of the means to identify the additional observation, as opposed to having both means. Presumably there is also potential in skewness and kurtosis statistics to generate a differencing as you have an extra equation now to consider – it might be you don't need one of the means, for example.

Riskiness

So, there clearly is potential risk here. How meaningful is it?

It is possible to conceive of a case where a researcher checks for variance, finds another observation and redoes it, and then reports both variances along with one of the means (if both means are reported, the differencing comes from them and the variances are irrelevant). This seems an unlikely occurrence; if it did happen, a disclosure would then require someone to realise that this is the case, calculate the additional value, and associate it with any characteristics of the additional observation which would allow for identification. Moreover, researchers are generally warned about this in training. Hence, we treat this as not meaningful.

On this basis the SD (and hence variance, skewness and kurtosis) can be classified as **safe**.

Assessment

Manual and automatic:

- Check residual degrees of freedom

Non-linear concentration ratios

Analysis

A Herfindahl index is a measure of concentration used in economic analysis. It is calculated as the sum of squared shares (where x_i is the value and T=total):

$$H = \sum \left(\frac{x_i}{T}\right)^2$$

The value varies between full concentration at 1 (x_1=T, x_2..x_T=0) and even distribution (x_1=x_2..=x_N) at 1/N. There is no differencing risk, but there is a potential dominance risk. If x_2..x_N are all close to zero then $\sqrt{H} \approx x_1/T$; similarly if there are two very large observations, and the rest are zero. These are stronger than for a simple dominance check on a total, because the squaring exacerbates the dominance.

Because of the squaring I haven't been able to come up with simple rule for dominance yet, other than the obvious "check root of H is not equal to the largest observation", which is the current guidance.

Riskiness

I would retain this as a **safe** statistic, as the dominance check ("check $\sqrt{H}$ is not close to the largest share") is easy to carry out, and once that has been confirmed there is no meaningful risk irrespective of the data. In discussing 'safe statistics', we also assume the statistic is likely to be safe even if the checks are not carried out, which seems reasonable in this case: dominance problems are not obvious to the reader, and require additional assumptions about any attacker. Moreover, indexes where H approximates to 1 are of limited research interest.

Assessment

Manual:

- Check number of observations
- If feasible, check whether largest share is within h% of the index

Automatic:

- Check whether largest share is within h% of the root of the index

where h% is determined by the data owner. Note that this is not he same percentage usedin the standard N,K and P rules.

Gini Coefficient

Analysis

The Gini coefficient is a measure of inequality from 0 (perfect equality) to 1 (perfect inequality). The discrete formula is (according to Wikipedia)

$$G = \frac{2\sum_i ix_i}{n\sum_i x_i} - \frac{n+1}{n}$$

This is potential identifiable if you have n=2 and one of the values, but otherwise, there seems little to worry about. It can't be differenced, because a new observation would alter the ranking, unless the new observation was larger than any other in the dataset (not entirely silly, as you might want to see the effect of adding a very rich person to a mix, for example in teaching). But even then I think you would need to do a lot of manipulation (will try calculations later).

Riskiness

This goes firmly in the **safe** box, at least for now.

Assessment

Manual and automatic

- Is n>2?

Smoothed distributions/modelled functions

These kernel density estimates for plots. These come out of models. We therefore roll them into regressions and drop as a category.

Small numbers

Many (perhaps all) of the statistics can be problematic if there is both a very small number of observations and a limited range of possible values for the source data. Derrick et al (2022) demonstrated this in the case of Likert scales. It is possible to conceive even of something like the observations underlying a regression being discoverable if there were only a few observations, one or two explanatory variables, all the variables could only have a few values, and the regression statistics were published with sufficient decimal places.

It's also clear that these are exceptional cases which we would not expect in genuine research output (with the exception perhaps of Likert scales in small studies), and they do require both awareness and incentive for someone to try to unpick this. Hence we do not think these are meaningful concerns.

However, this potentially provides a useful way to think about why thresholds above 2 might be a good idea. So we will be working on this (simulations) but not directly feeding into SACRO in the short term.

Dominance

We don’t formally consider dominance in the more complex statistics (SHTs, regressions, shape). It seems that this is a theoretical possibility but some basic analysis in Ritchie (2014) and by, I think, Statistics New Zealand suggests you really need such an unlikely dataset that we can safely discount it occurring in genuine published outputs.

Appendix 3 Class disclosure, and evidential and structural zeros

A zero in a table can lead to class disclosure. Consider the following (adapted from ONS, 2019):

Highest education level	Income quartile				
	Q1	Q2	Q3	Q4	Total
University	2	4	27	34	67
College	15	28	76	51	170
School	13	16	22	0	51
None	8	12	0	0	20

Table 35 Example of class disclosure through an informative zero

The zeros are **informative**: they indicate that no-one with just school qualifications is in the highest earnings quartile, and that no-one who left school without qualifications earns above the median income. These are **class disclosures**: knowing that someone is a member of the class allows new information to be determined about them.

Not all zeros are informative. *Structural* zeros by construction must be zero. For example (taken from the DRAGoN output checking course), the table shows numbers of young people in the UK Labour Force Survey earning above and below the National Minimum Wage:

Age last birthday	Paid below NMW	Paid at or above NMW	Total
16	0	1366	**1366**
17	0	1258	**1258**
18	114	990	**1104**
19	63	1003	**1066**
	177	**4617**	**4794**

Table 36 Example of no class disclosure due to structural zeros

This example is taken from 2002, when there was no minimum wage for those aged under 18. Therefore, the zeros are not informative, but **structural**: they are zero because of the definition of the categories. If a non-zero value is found in a structural zero, this indicates an error in the data.

The SDC literature generally only distinguishes between structural and non-structural zeros. However, we identify a third category, that of **evidential** zero. This is a cell which is not zero by definition, but we would expect it to be empty. Two examples:

	Highest qualification			
Age	None	School	College	University
16-17	12	84	0	0
18-20	8	33	62	0
21-25	9	18	21	37
26-30	14	21	17	42

(a) All evidential zeros

Cancer	Males	Females	Total
Bowel	321	423	743
Breast	0	109	109
Lung	51	23	74
Skin	18	9	27
Cervical	0	43	43

(b) Evidential and structural zeros

Table 37 No class disclosure due to evidential zeros

The green boxes in both tables are evidential zeros. It is possible that a 16 year-old could have a degree, and males have a very low but nonzero likelihood of breast cancer; but in both cases we would not be surprised to see these categories empty. On the contrary, a positive value in these cells would be highly disclosive eg a ‘child genius’ who gets her degree at Oxford aged 13, widely reported in the press.

The red box in table (b) is a structural zero. This shows that it is possible to have both types of zero in the same table.

Appendix 4 Full statbarn listing

This is an overview of the statbarns as currently defined.

Statbarn	**Description**	**Risk**	**Things to look for**	**Mitigation strategies**
Frequencies	This class covers frequencies ie counts of things, either in tables (most common), in certain graphs such as histograms or bar charts, or single as in a description of the number of survey participants. This also includes frequencies expressed as a proportion of some total	Unsafe (High risk)	Many issues need to be considered to accurately assess the risk of any individual table. For example, is the data itself disclosive? Could units making up the data or subsets be identified? Is the rank ordering of contributors known? Or, what is the sample choice/weighting/cell units? Low counts (typically counts below ten are considered at risk of being disclosive), Differencing between released tables of a dataset, Class disclosure. Level of geographic disaggregation. Detail of industrial/Occupational classification. Global context.	Suppression can be used to remove cells that fall below a threshold. Rounding can obscure the true value of a cell to protect against differencing. Noise addition can also be used to protect a table.
Statistical hypothesis tests	Statistical tests are used to make inferences and observe differences in data. These involve a wide range of tests including t-tests, p-values, F-tests, or confidence intervals.	Safe	Residual degrees of freedom (broadly, number of observations less number of restrictions implied by the tests)	N/A - No meaningful mitigation
Correlation coefficients	Correlation coefficients are statistical measures that quantify the relationship between two or more variables. This includes measures such as regression estimates, as well as single statistics such as Pearson's r, Spearmans's rank correlation coefficient (ρ) and Kendall's rank correlation coefficient (τ).	Safe	Theoretical risks include saturation (all values of all variables fully interacted), or no residual degrees of freedom ie an equation rather than a regression. Only the latter is meaningful, and only when carrying out a regression with one or two binary variables. If this case holds, the 'regression' should be treated as a one-way or two-way table of means.	N/A - No meaningful mitigation

Position	Position refers to a statistic that provides information on a central or typical value in a dataset. These include data points such as median, percentiles, or inter-quartile range	Unsafe (Low risk)	Percentiles should be treated as special cases of magnitude tables, each percentile band should be treated as a tabular cell with population determined by position in the rank. To decide if this value is disclosive, the size of the cell and range of the band must be considered. Low counts and class disclosure are the main issues with this type of statistic. Understanding whther the position of an observation in the ranking is crucial to determining whether there is any disclosure risk or not.	Rounding or noise addition can be used to obscure the underlying values to make a positions statistic safe for release. Suppression can also be used in some cases. Removal of outliers can be helpful in preventing percentile values for being skewed in a way that may make data vulnerable.
Shape	Shape outputs refer to measures that describe or show the characteristics of data distributions. They provide information on elements such as symmetry, peaks and tail behaviour. These comprise standard deviation, skewness, and kurtosis.	Safe	Low degrees of freedom can prove an issue for disclosure in shape outputs.	N/A - No meaningful mitigation
Linear aggregations	This class covers sums, means, ratios etc. These are stats that provide a snapshot of the data's characteristics that don't relate to any one data point and instead are calculated with many or all of the points in a data set. This class also includes linear concentration ratios (eg share of the top N observations).	Unsafe (Medium risk)	Low counts can be an issue with linear aggregations. Differencing can in some cases be applied to linear aggregations to expose underlying data. Dominance can also impact these statistics and can cause issues for data protection.	For each value to be released, the largest contributor included in the synthesis cannot exceed a proportion of the total (proportion set by data owner). Depending on the nature of the risk, the statistic could benefit from suppression, rounding, noise addition and/or outlier removal.
Mode	The value or values that occur most frequently in a dataset. It represents the peak or highest point on a distribution's histogram.	Safe	If all units have the same value then this would count as a class disclosure and should be assessed as such	N/A

End points	Minimum and maximum values for a variable	Unsafe (High risk)	The threshold rule may be broken by maxima and minima referring to a single data point (more likely to happen with a maximum on a variable of unrestricted range). They can also risk class disclosure as they provide information on every value in the dataset. Structural maxima and minima (eg a percentage ranging from 0% to 100%) are less likely to be problematic.	Suppression is most likely to be the only solution. If a maximum or minimum is far enough out to be problematic, it is unlikely that rounding or noise addition can make a significant difference. However, capping values can be very effective.
Non-linear concentration ratios	Non-linear concentration ratios are statistical measures used to assess degree of concentration or dispersion. The most common of these are the Herfindahl–Hirschman index.	Safe	There is a dominance issue if the distribution comprises one large value and every other value negligible. This is unlikely to be meaningful in practice	N/A - No meaningful mitigation
Calculated ratios	Calculated ratios refer to measures investigating relationship, comparison or relative magnitude between variables. These can includes odds ratios, risk ratios, or hazard ratios.	Unsafe (Low risk)	While the ratios themselves are safe, the likelihood of publication as some marginal totals (for example, numbers in treatment and control groups) means that the underlying contingency tables can be recreated. The underlying tables need to be published	Rounding may be the most appropriate for small numbers, or reducing precision for many observations.
Hazard/survival tables	These are tables used to analyse and model survival or failure times in event-based data and business data. Commonly seen in epidemiology, actuarial science and clinical research among others. These can also underpin visualisations such as Kaplan-Meier.	Unsafe (Low risk)	Generally very low risk, but be careful when using absolute dates or when extreme outliers are presented	Convert absolute dates to relative dates; right-censor extreme observations
Linked/multi-level tables	Multilevel and linked tables are used to summarize and analyze the relationship between multiple categorical variables. These often analyse nested categorical data in formats such as two or three way contingency tables or hierarchical contingency tables.	Unsafe (High risk)	Unknown.	Unknown.

Clusters	Cluster analysis is a statistical technique used to classify or group similar objects or individuals based on their characteristics or attributes. It aims to identify patterns, similarities, or relationships within a dataset by grouping data points into clusters.	Unsafe (High risk)	Unknown.	Unknown.
Gini coefficient	The Gini coefficient is used to summarise inequality within a population. The Lorenz curve is its graphical counterpart.	Safe	Only problematic if trying to calculate from a population of two…	N/A - No meaningful mitigation