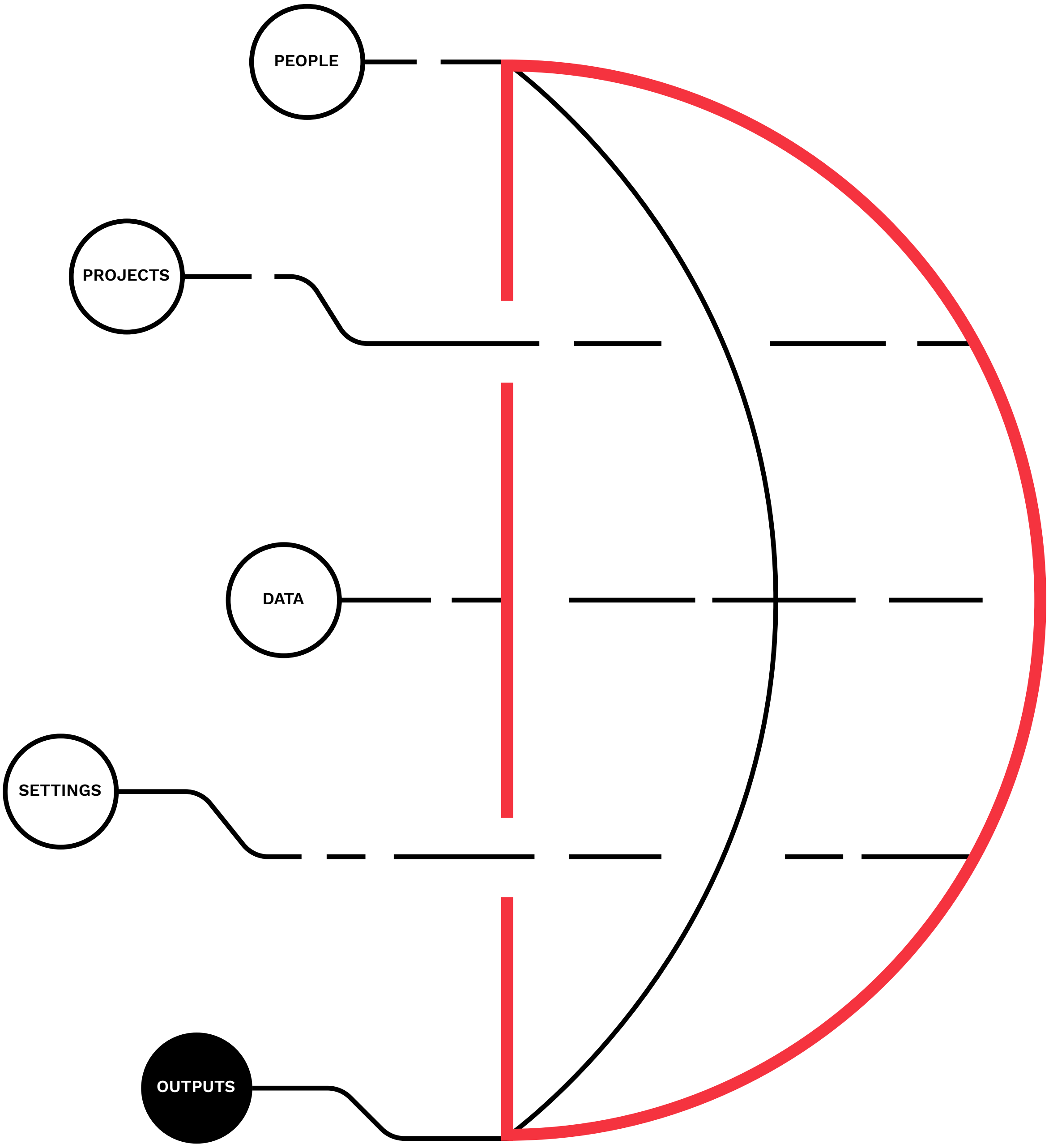
Handbook on Statistical Disclosure Control for Outputs

Emily Griffiths (University of Manchester)
Carlotta Greci (The Health Foundation)
Yannis Kotrotsios (Cancer Research UK)
Simon Parker (Cancer Research UK)
James Scott (UK Data Archive, University of Essex)
Richard Welpton (The Health Foundation)
Arne Wolters (The Health Foundation)
Christine Woods (UK Data Archive, University of Essex)



July 2019 Version 1.0

Produced by the Safe Data Access Professionals Working Group

Acknowledgements

The authors wish to thank the following individuals for their guidance and support, and for reviewing this document:

Anthea Springbett, **Felix Ritchie** (University of the West of England), **Mark Elliot** (University of Manchester), **Matthew Woollard** (UK Data Archive, University of Essex), and **participants from the August 15th 2018 SDAP workshop on SDC** (including staff from HMRC Datalab, ONS Secure Research Service, UK Data Service, Cancer Research UK, NHS National Services Scotland, The Health Foundation).

The authors are grateful for the kind and generous support of The Health Foundation for supporting the publication of this Handbook.

About this version

Welcome to version 1.0 of our SDC Handbook. Following the release of the beta version in March 2019, we're pleased to hear that staff undertaking disclosure checks are using the guide and finding it helpful for their day-to-day jobs.

This reflects two facts:

- in recent years, many new secure data environments (Safe Settings) have been established across government, academia and elsewhere, and have recruited staff tasked with the responsibility for ensuring that statistical results produced from confidential data pose a minimal risk of disclosure of identity and/or personal information;

- the staff recruited believed they had little in the way of resources to turn to, and so as they reviewed drafts of this Handbook, they began using it in their roles.

To bring this Handbook about, the authors have pooled their collective knowledge of SDC and explained how they undertake SDC assessments. However, we acknowledge that there is more work to be done, and that subsequent versions should be updated. There are others who undertake these roles who we haven't spoken to and who could contribute their own examples. There are researchers in statistical confidentiality and privacy who we would be grateful to hear from, particularly to ensure the technical accuracy of the contents. We'd like to hear from anybody who has ideas about how this Handbook could be improved. Please do contact us by visiting https://securedatagroup.org/sdc-handbook.

About the Safe Data Access Professionals group

The Working Group for Safe Data Access Professionals (SDAP) was established in 2011 to bring together staff working in Safe Settings to share experiences and develop best practice.

The network is made up of members working to provide secure access to confidential business, health and socio-economic data, in government, academia and charities. This piece of work is part of a wider engagement from the SDAP group in sharing best practices for, and expertise in, managing access to sensitive data.

More information about SDAP can be found at securedatagroup.org.

Responsibility for the information and views set out in this Handbook lies entirely with the authors and does not necessarily represent the views of the organisations they work for.

Contents

Introduction 6

Audience 9

Developing this Handbook 10

Scope 12

Aims 12

Structure 13

Key concepts 14

SECTION A **15**

The statistical risk: what is it all about? 16

What is Statistical Disclosure Control? 19

Risk assessment 22

The statistical risk: principles and rules 26

SECTION B **28**

Introduction: SDC 29

Using this guide 30

Descriptive statistics 31

Percentiles 33

Histograms 35

Box plots 38

Correlation coefficients 40

Factor Analysis 42

Indices 44

Scatter plots 46

Symmetry plots 48

Decision trees and exclusion criteria 51

Survival analysis: Kaplan-Meier curve 53

Spatial analysis (maps) 55

Gini coefficients 57

Concentration ratios 59

Regressions 61

Residuals 64

Margin plots 66

Test statistics 68

SECTION C **69**

Introduction 70

Implementing SDC as an organisation 71

Managing analysts 76

What is a good set of statistical results? 78

The good 81

The bad 82

Managing expectations 83

Further resources 85

Glossary 88

Copyright 90

Introduction

Analysts are demanding access to more data about individuals and organisations than ever before. By using more detailed data, it is possible to do more robust and innovative analyses, explore new themes and strands, and generate results that better support policymakers and business decisions.

Such data sources are now routinely available. However, the level of detail is such that these data are considered 'personal, sensitive or confidential' and are subject to Data Protection laws. As such, access to these data is typically made available in a 'Safe Setting' (a secure facility to access sensitive data) to ensure the confidentiality of the data subjects is preserved.

ABOUT THE FIVE SAFES

The Five Safes is a way of thinking about how to access data. It was originally devised by Professor Felix Ritchie (now of University of West of England) during his time at the Office for National Statistics (around 2006).

The framework provides a decision-making process to enable 'safe use' of personal, confidential data. 'Statistical purposes' were what was in mind when the framework was drawn up.

Specifically, the framework considers that there are five aspects of safe data access:

- **Safe Data** (what are the characteristics of the data? Are they sufficiently detailed that they contain confidential information attributable to individuals?)

- **Safe People** (who is going to access the data? Do they have the right credentials, experience and motivation for accessing data?)

- **Safe Projects** (what task is going to be undertaken with the data? Is it research, or does the person accessing the data want to try to identify somebody? Or is there another reason for accessing the data? Is there a 'public benefit'?)

- **Safe Settings** (consider the environment in which the data will be accessed? Are there safeguards in place to help protect the confidentiality of the data?)

- **Safe Outputs** (what will be released? Statistics? Can these be used to identify somebody? Could confidential information be released?)

This is a framework widely used, not just at the Office for National Statistics, but other statistical agencies around the world (including Australia and Nepal), as well as organisations such as the UK Data Service, Cancer Research UK and The Health Foundation. More information about how the Five Safes framework can be applied is to be found in Desai, Ritchie and Welpton (2016): see Further Resources.

This Handbook has been designed specifically for people who want to ensure that Safe Outputs are produced from personal and confidential sources of data (the fifth of the Five Safes listed above). See Audience for more information.

SAFE SETTINGS AND SAFE OUTPUTS

In this framework, analysts can use the Safe Setting to access and analyse sensitive data. The statistical results generated in the Safe Setting are retrieved only after they undergo a review of disclosure risk (i.e. Statistical Disclosure Control, SDC) to ensure that the published results do not reveal the identity or contain any confidential information about a data subject (an observation in the data). Safe Outputs are only released from the Safe Setting.

A number of Safe Settings operate throughout the UK. This includes government agencies such as the Office for National Statistics' Secure Research Service and HM Revenue and Customs' Datalab; academic-funded infrastructures such as the UK Data Service's Secure Lab; and charities such as The Health Foundation and Cancer Research UK. Recently, Connected Health Cities, a joint academic-health sector partnership, established a network of Trustworthy Research Environments to analyse patient data. These organisations have established their own Safe Settings to acquire and provide access to confidential data for use by their analysts or to provide access to third-party analysts. Within this framework, it is paramount to ensure that any statistical results that are published do not pose a risk to the privacy of the data subjects behind the data. SDC plays a critical role in mitigating this risk and represents a key component of information security systems of Safe Settings.

This Handbook aims to achieve two broad objectives:

- introduce the principles of SDC with an overview of the major challenges and best practices in the UK;

- support organisations and staff to practically apply SDC to statistical results generated from confidential data.

To some extent, this work is also intended as a source of assurance for organisations engaging with data owners and data subjects. Following the implementation of the 2018 General Data Protection Regulation (GDPR), more organisations have an incentive to gain information security accreditation (such as the ISO 27001 standard or Data Security and Protection Toolkit). It is now more important than ever that Safe Settings can prepare their staff to release safe statistical results. This Handbook provides organisations with guidance about how statistical results generated from sensitive data are released, whilst ensuring the safety of the data.

Audience

The Handbook has been written for two audiences:

- staff working in Safe Settings who are responsible for applying SDC to statistical results derived from confidential data;
- analysts who are planning to publish findings created from data held in a Safe Setting.

For either audience, this Handbook assumes the reader has familiarity with basic statistical concepts and experience of handling data. The SDAP Competency Framework (securedatagroup.org/guides-and-resources) provides guidelines about how new staff can develop their skills, which includes SDC.

It is the responsibility of the organisation hosting the Safe Setting to train their analysts on how to check statistical results prior to release. However, we intend this Handbook to be a useful reference resource and, ultimately, to help achieve a consistent approach in the SDC assessment of statistical outputs generated from confidential data. Each example of statistical outputs presented in this Handbook comes with a set of 'would be useful to know' items that can be used in any SDC system and by any type of user.

Developing this handbook

In April 2017, the authors, together with staff from the ONS, met to discuss and compare requirements and guidelines for staff checking statistical results to ensure they were safe to be released. One feature of the group is that staff working at SDAP member organisations undertake SDC in some way. Undertaking SDC is generally a key component of managing a Safe Setting.

SDAP members have generally followed the ESSnet/DwB publication, “Guidelines for Checking of Outputs”; and, in health, the “ISB1523 Anonymisation Standard”, to undertake SDC. The Information Commissioners Office (ICO) has published an “Anonymisation Code of Practice”. However, the group considered these guidelines would benefit from improvements, given the advances in statistical methods, and more widespread use of confidential data in Safe Settings.

The ESSnet guidelines are one of only a small number of sources of information which the SDAP group is aware of, that solely provides specific advice about assessing statistical results for disclosure risk (although there is a large literature on statistical privacy, this tends to be aimed at publishers of national statistics). Analysts nowadays produce a larger variety of statistical results than tables, such as maps, descriptive statistics, modelled outputs. Safe Settings should be able to release any type of output that analysts produce, and ensure that these outputs will not breach confidentiality of the data when released. This flexibility therefore requires staff responsible for checking and releasing statistical results to have sufficient experience in applying the principles of SDC to a variety of analysis types.

Much academic literature has been published on the topic of statistical disclosure control. For example, one could browse articles in the Journal of Privacy and Confidentiality or Transactions in Data Privacy for technical advice on statistical disclosure control. SDAP believe that this literature provides useful concepts; however, it is vital that lessons can be practically adopted and assessed.

With this in mind, SDAP agreed that it would be useful to develop an SDC Handbook to use:

- for practical guidance about how to assess most common SDC requests while providing a basis for assessing more complex cases;
- for advice on how to encourage the production of good and acceptable statistical results, the release of which can be facilitated by the SDC process.

This will help to ensure:

- that statistical results are assessed accurately and comprehensively;
- that statistical results are checked consistently and are, as much as possible, aligned with other Safe Settings practices.

The authors and SDAP intend that this Handbook will be continuously reviewed, to ensure that it is useful, up-to-date, and serves the needs of individuals assessing statistics in Safe Settings; data suppliers, who require assurance that the confidentiality of the data they make available is not compromised; and those developing and delivering training courses for both staff and external analysts.

FUTURE UPDATES

As this is a work in progress, this Handbook will be regularly updated. securedatagroup.org/sdc-handbook

SCOPE

This Handbook is about the assessment, management and release of statistical results produced from confidential data sources held in a Safe Setting. By ‘statistical results’, we mean statistics derived from the data and not other outputs that could also be produced, such as:

- derived datasets (e.g. subsets of the data, or aggregated datasets);
- synthetic versions of confidential data sources;
- weights derived from confidential data sources.

AIMS

To summarise, the aims of this Handbook are:

- to translate disclosure control concepts into practical advice, measures and steps for assessing statistical results for disclosure risk;
- to assure data owners that access to data supplied by them is securely managed, and that data confidentiality will not be compromised;
- to make the process of requesting releases of statistical results easier;
- to be used with training for staff working in Safe Settings with examples that can be used directly for training purposes.

We think this Handbook will provide a robust reference source for staff working in Safe Settings to gain information governance accreditation, in light of the development of new statistical techniques, data sources being shared and combined from different disciplines (such as health and social sciences), and changes in the legal and regulatory environment.

Structure

THIS DOCUMENT HAS THREE SECTIONS.

Section A, for those looking for an introduction about how to make statistical findings ‘safe’. This could include analysts checking their outputs, and staff undertaking assessments.

Section B provides practical step-by-step guidance for undertaking disclosure control assessments.

Section C is for staff setting up and managing a data ‘service’ where Statistical Disclosure Control is a significant contribution to staff roles.

Sections A and B provide an overview of Statistical Disclosure Control and defines the concept of disclosure risk. An extensive list of common statistical outputs are described, with accompanying examples and practical guidance about how to undertake SDC. These include frequency tables, regression coefficients, histograms, plots of residuals, and more complex techniques such as survival and spatial analysis. For each type of output, advice is also provided about what to look out for, and we suggest questions that could be asked of the individual submitting and/or assessing the statistical results for release. We recognise that there are other types of statistical results that may need to be assessed for statistical disclosure risk; guidance is also provided about how to manage statistical disclosure of new types of statistical results never previously assessed.

Section C provides advice and guidance for organisations about how to manage the SDC process and workload. We also consider other methods of reducing the risk of releasing statistical results with the potential to breach data confidentiality, including how to incentivise good output requests.

Key concepts

Throughout this book, we refer to a number of terms. We've listed some here for guidance, they can also be found in the Glossary.

OUTPUT CHECKERS

Those responsible for checking statistical outputs (a.k.a. statistical results) created in Safe Settings for potentially disclosive issues.

SDC

The process applied to statistical outputs (statistical results) to mitigate the risk of potentially disclosive results leaving the Safe Setting.

SAFE SETTING

The technical means – whether physically located or virtual – through which the analyst works on the data.

DATA SUBJECT

The unit of observation in a dataset. Usually individuals or businesses, depending on the source of the data.

SECTION A

About Statistical Disclosure Control

The statistical risk: what is it all about?

IN THIS SECTION YOU WILL LEARN ABOUT:

1. The concept behind how somebody could be identified from published statistics
2. How published statistics could breach confidentiality

THE CONCEPT BEHIND STATISTICAL DISCLOSURE CONTROL

The basic idea is that, using some statistical information, it could be possible to infer confidential information, and even identify someone, from a set of results that have been released.

Supposing we have some data about a bunch of companies operating in different sectors. The data we have includes turnover and employment for each company. How might statistical disclosure and identification occur?

Let's start by aggregating the turnover data by sector. We'll also know the number of companies in each sector.

We might find that for one sector, there are two companies. We could publish the turnover figure for the sector, which is the sum of both companies' turnover.

However, if we did so, then one company, seeing the results, could work out the turnover of its rival. This is because it knows its own turnover, which it could deduct from the published aggregate figure, and therefore associate the remainder with its rival.

For example, if the turnover for the sector was £1,000,000, and Company A knew its turnover was £400,000, then if there are only two companies in the sector, it must be the case that the remaining £600,000 is the turnover of Company B, its rival.

Now suppose we add a third company to the mix, Company C. So we now have a published aggregate figure of £1,200,000 for three companies in the sector. Although Company A still knows its share is £400,000, it can only ascertain that the remaining £800,000 is shared somehow between Company B and Company C. But without knowing more, it cannot exactly apportion the figure between Company B and Company C, and, therefore, statistical disclosure hasn't occurred.

This explains why, from a statistical perspective, there should always be a 'threshold' of at least three observations, i.e. the minimum frequency that a statistic is based upon. However, Safe Settings often use a higher threshold than this, to account for 'secondary disclosure'. This is explained in the following two sections.

STATISTICAL CONFIDENTIALITY BREACHES

In the previous example, we illustrated how statistical disclosure could occur using an example from business data (given companies over a certain size are obliged to publish accounts, you might wonder why we used this example: the ONS provides secure access to many data sources about businesses which are actually collected in confidence via survey pledges).

Here is another example.

Cancer statistics

A number of health agencies regularly publish statistics on cancer incidence. Sadly, cancer isn't uncommon, so one might expect that a breakdown of cancer incidence by local authority to be fairly numerous. This isn't always the case. For example, the City of London, while well-populated by financial institutions with many workers, does have some residents. Therefore even publishing incidence of a common cancer, such as lung cancer, would probably display as small frequencies. For rarer cancers, the numbers would likely be in single figures, and probably one would find counts close to 1 or 2. If you were the person diagnosed with the illness, you might be able to recognise yourself in the statistics. Others may recognise you too. There are two issues here:

1. you have been identified;

2. some confidential medical information about you has been released;

(and thirdly, the confidential medical information has been associated with you).

DISPUTES

Department of Health, Abortion statistics, 2011

Sometimes making a decision about whether to release statistics can be difficult. The 2011 case involving the Department of Health, which decided not to publish statistics relating to abortions over 24 weeks, is a case in point.

When broken down by certain categories, the frequencies were in single figures. In one case, it was possible to ascertain that one abortion was due to a probable disfigurement of the child. The numbers were so small, the Department of Health argued against publishing on the grounds that the women could be identified. A Freedom of Information (FOI) request demanded that the statistics was submitted, and when refused, a legal case ensued.

Eventually the High Court ruled that the statistics should be published, because as statistics, the personal data were rendered anonymous and not personal data. This highlights the issue about when published statistics can be reversed back into personal data.

We include this example to explain that SDC isn't always a cut-and-dry topic. We have written this guide as a risk-assessment approach. The fact is that it might be entirely safe to release a statistic consisting of small frequencies: nobody will be identified or harmed. On the other hand, it might be in the public interest to release statistics even if disclosure is likely to occur. However, output checkers are not always responsible for making these decisions; the data controller may have to decide this.

FURTHER READING

1. ICO Anonymisation Code of Practice, pp 14-15
   https://ico.org.uk/media/1061/anonymisation-code.pdf

What is Statistical Disclosure Control

IN THIS SECTION YOU WILL LEARN ABOUT:

1. Statistical Disclosure Control
2. How this Handbook approaches SDC
3. Why the source of the data matters
4. The importance of the unit of observation

Statistical Disclosure Control (SDC) aims to:

- prevent the identity of a data subject from being revealed;
- and/or releasing associated confidential information belonging to that data subject.

Traditionally, SDC has been applied to:

- statistical tables (often produced by National Statistical Institutes) prior to their release;
- microdata, for the purposes of creating anonymised versions of original data.

However, the rise in the number of Safe Settings and the possibility of complex analyses has led to further development in techniques to assess disclosure risk from statistical outputs. After all, there is little to be achieved in establishing a Safe Setting with secure access to confidential data if the statistical results that are released breach the confidentiality of the data.

The risk of re-identifying a data subject (what the data are about, e.g. a person, organisation) from these types of data sources is a real one and needs to be managed carefully and efficiently. In this Handbook, SDC relates to the assessment of statistical results produced from confidential data, rather than statistical tables and microdata for public release.

Note that for the purposes of this Handbook, we make no distinction between whether individuals for which the data are about are living or dead (for that matter, whether the companies which business data are about are solvent or bankrupt/closed). While the Data Protection Act covers living persons, the Census Act contains a 100-year clause to prevent publication of individual's details; and the Statistics of Trade Act does not distinguish between live or 'dead' companies. To keep things simple, we assume that we should protect the confidentiality of data and the identities of all data subjects in the data.

Although SDC aims to prevent the re-identification of a data subject, this is a 'risk minimisation' strategy rather than a 'risk elimination' one. Few would wholeheartedly claim to have removed disclosure risk entirely, particularly when the context by which personal data remain confidential may change. For example, people may publish their details on social media while they may have previously expected such information to remain confidential. Trying to eliminate the risk of statistical disclosure entirely would ignore changes in how we think about our data and our appetite for disclosing confidential information about ourselves; and we may ultimately fail in SDC as a result. Instead, risk minimisation takes account of outside conditions that we cannot control.

The approach presented in this Handbook brings together these considerations: the risks of disclosure should not be ignored, but instead managed and mitigated as far as possible. The only way to achieve risk elimination is not to release any statistical results from a Safe Setting: but such a course of action would be detrimental for analysis, research and the public good in which they serve. A balance is struck when statistical outputs can be released while ensuring the risk to confidentiality is minimised as much as possible, and that due diligence is observed.

APPROACHES

A Safe Setting facilitates access to confidential data for analysis. Since an analyst could devise any type of result from their analysis, Safe Settings need to be prepared to assess any type of statistical output. Any statistical result should be assessed on whether there is any risk that a data subject could be re-identified, or whether confidential information could be revealed.

In some cases, the public benefit from releasing a statistic may be considered to outweigh any such risk. For example, in a clinical trial it could be beneficial to publish the fact that one person had a severe reaction to a new treatment, if the conclusion from this result is that a drug is withdrawn pending further development and testing. One person might be singled out and indirectly identified, but the benefit to others could be considered to exceed this risk.

DOES THE SOURCE OF THE DATA MATTER?

The source of the data is relevant when making an SDC assessment. For instance, it could be argued that survey data contain a lower risk of identifying any data subject because of the nature of the sample (and the sampling framework), whereas administrative data include everyone relevant, so there is more certainty that a possible identification is accurate. When making an SDC decision, whether the data originate from a source or administrative data can be factored into the risk assessment.

UNIT OF OBSERVATION

Care should be taken to understand the unit of observation or data subject whose confidentiality we are trying to protect.
SDC can apply to individuals, businesses, households, or particular sub-groups of population such as patients, taxpayers or vulnerable people. While the application of SDC follows the same process, the context and meaning of disclosure associated with that unit of observation may vary.

For example, an analyst of a business dataset may have produced statistics based on sites (including low level geography such as postcodes). If sites (e.g. supermarket branches) belong to the same organisation, then the risk of re-identification of the organisation may prevail, even if, for example, all the statistics are aggregated and meet frequency requirements for associated individuals. At a glance, it might appear that the statistical results are able to be released; but if the unit of observation is not what it first might seem, then extra care should be taken to consider the context of the analysis and the disclosure implications.

FURTHER READING

1. Duncan, G., Elliot, M., Salazar-Gonzalez, J., (2011) “Statistical Confidentiality: Principles and Practice”, published by Springer

Risk assessment

IN THIS SECTION YOU WILL LEARN ABOUT:

1. Identification, attribution and secondary disclosure
2. The importance of contextual information
3. Whether to check all results produced by a user or not
4. Types of risk that data suppliers are concerned about
5. Concepts of thresholds and dominance

KEY CONCEPTS

IDENTIFICATION

This occurs when a data subject is identified from statistical results. SDC aims to minimise the risk of this happening.
Identification can occur when a small number of observations are isolated and presented in an output. For example, calculating the maximum income from a dataset will probably reveal the exact income of one person. When combined with other statistical results such as age or location, the person with the highest income could be identified and their confidential data revealed.

ATTRIBUTION

When a number of characteristics can be put together and associated with the same observation (even in the absence of direct identifiers such as name and address), a data subject might be identified. Disclosure by attribution may occur when characteristics, seemingly anonymous individually, are fitted together. As with identification, this may occur when a small number of observations are presented in a statistical result.

For example, a scatter plot displaying the values for two variables for each data subject will probably enable information to be associated with a data subject. Additionally, there are other scenarios in which attribution can occur. For instance, if a group of data subjects share one characteristic, presented as a statistical result, that characteristic can be attributed to each data subject.

SECONDARY DISCLOSURE

Secondary disclosure occurs when one released result is combined with other information to produce new statistics that are disclosive.

For example, two tables of descriptive statistics produced from one data source could show various breakdowns in different ways. A single observation could be isolated, for example, if one table is deducted from another.

Alternatively, statistics could be combined with data available outside the Safe Setting where the data were analysed. There is always a potential risk that even 'disclosure controlled' statistical results could be combined with other data or results to reveal the identity of a data subject.

CONTEXTUAL INFORMATION

Assessing the risk of a statistical result can be aided significantly when relevant contextual information is available.
This could include, but is certainly not limited to:

- whether a specific industry or occupation is being examined;
- whether the data or results have been pooled and averaged over a specific time period or geographic area;
- whether there is a particular knowledge of the population underlying the data;
- underlying unweighted sample sizes for different population sub-groups in more than one table or graph;
- other, likely, available information already in the public domain or available to those who will receive the results.

Without context or information about the statistical results there is a higher degree of uncertainty regarding the disclosure risk from the statistical output. Amidst this uncertainty statistical findings should not be released from the Safe Setting unless adequate information is provided.

IS IT NECESSARY TO CHECK EVERY RESULT?

Analysts tend to produce more statistical results than those requested for release and published. As methods are revised and analysis follows its iterative path, the analyst will decide which statistical results are worth requesting for release.

Within the Safe Setting framework, every request to release a statistical result must be checked through SDC. Some results will be more demanding to assess in terms of time or complexity than others. For example, results from a regression analysis will take less time to assess than a table of descriptive statistics. These differences will be explained in more detail later in this Handbook. Nevertheless, all requested releases should still undergo SDC. Imagine the consequences that would follow if an output was not checked before released and revealed the identity of a data subject, causing harm and distress to that individual, and bringing penalties for breaching the law.

THRESHOLDS

Most disclosure risk assessments judge statistical outputs on a set 'threshold', defined as the minimum number of observations underpinning the statistics.

In a table of frequencies this is straightforward: if a cell has a frequency value less than the required threshold then the table is deemed 'unsafe' for release, unless it is proven after further scrutiny with contextual information that it does not lead to the identification of any specific data subject.

This said, it is important to note that tables cannot be judged in isolation. As explained previously, secondary disclosure could occur when two tables are differenced, even if the threshold is met in each table individually.

Many data owners have different requirements for the value of the threshold. This ranges from anything from a minimum of three to a maximum of 30 data subjects. This reflects the risk appetite of the data owner and the nature of the data.

Throughout this Handbook, it is assumed that the threshold will be set to 10 (i.e. N=10). This value is used by the Office for National Statistics (ONS) for England and has subsequently been adopted by a number of other Safe Settings and government departments.

Sometimes it is possible that a statistical result is not disclosive even if the threshold has not been met. Conversely, the threshold rule may be met but the statistical result is unsafe to release. This will depend on the nature of the statistical result: a risk assessment of each result is a valuable exercise. Rather than using a more rigid 'rule-based' approach (i.e. the results cannot be released unless the threshold is met), we advocate a principles-based approach where each result is assessed for potential disclosure risk; and if it is deemed that this risk is negligible then the results may be released.

DOMINANCE

This is the idea that one observation could account for most of the value in a statistical measure, and therefore be identifiable. It can sometimes apply to individuals, but is more of a concern for business statistics where firms might dominate a market or sector, or might be making large investments in a particular year.

See 'Concentration Ratios' for a worked example.

RISK APPETITE

Safe Settings must assure data owners that their data remain confidential. This involves providing assurances that SDC practices mitigate various risks (outlined below). Different data owners have differing risk appetites, which is reflected, for example, in how statistical results are assessed for disclosure and the threshold value. This Handbook aims to set out guidelines for undertaking a risk assessment of statistical results and to maintain the confidentiality of the data as much as reasonably possible, whatever the risk appetite of a data owner. These guidelines are flexible, and the parameters within the guidelines can be adjusted to suit the particular risk appetite of any given data owner.

PUBLIC PERCEPTION AND REPUTATIONAL RISK

If confidential data are released, then public trust and the reputation of the data owner could be damaged. As well as legal penalties, the operation of the data owner (which could be a government department providing a service) could also be inhibited. A small frequency, risk-assured, may 'seem' risky to the public, who may question why the results were released, even if there is little or no risk of a breach of data confidentiality.

PRIVACY RISK

There is a risk that statistical results released from the Safe Setting could lead to the identification of a data subject. Confidential information about the data subject may be released too. This could lead to harm and distress experienced either by the data subject or by people closely associated with them. There could also be legal consequences.

The statistical risk: principles and rules

IN THIS SECTION YOU WILL LEARN ABOUT:

1. Rules-based vs principles-based SDC
2. How this Handbook approaches SDC

The 'statistical risk' is about the minimum statistical information required to reveal the identity of an individual data observation, and/or confidential information about them. SDC is about preventing this occurring, as much as it is reasonably possible to do so, while balancing privacy risks with the public benefit of releasing some statistical information generated from confidential data.

Broadly, there are two 'schools of thought' about undertaking SDC:

- rules-based SDC;
- principles-based SDC.

Many services adopt a 'rules-based' approach, which states that, given a set of fixed rules about what can and cannot be released, the statistical output presented by a user either meets the criteria (if so, decision is to release), or not (do not release). For example, if a 'threshold' for frequencies presented in tables is set, such as 'no cell in a table should contain frequencies of less than 10', then a rules-based regime would allow the output to be released if all frequencies met the threshold; it would not be released if any frequencies failed to meet the threshold.

The 'principles-based' method is to simply ask that, given an assessment of risk, is the statistical output presented 'safe' to release or not? (in accordance with the Five Safes 'Safe Output' element). It might be that in a table of frequencies presented by the user, some cells do not meet a minimum threshold; but by taking a number of factors into account, it is decided in a risk assessment that it is ok to release the table.

An example could include frequencies for cancer incidence by cancer type for the whole of the UK. Cell counts of 2 and 3 are often released because without any further breakdown of the results, and no other information, it's almost impossible to work out who those 2s and 3s relate to.

Conversely, there could be examples where, even if the threshold for frequencies is met, it could still be problematic to publish the results because of the risk of disclosure.

For more information, Ritchie and Elliot (2015) have written on this topic (see Further Reading).

What are the advantages of the two regimes? The rules-based approach can be followed simply, often without thought. On the other hand, sometimes statistical outputs will not be released even though it is safe to do so, and this would be detrimental to research efforts. Moreover, and as implied above, it can provide a false sense of security, leading to unsafe statistics being released just because the rule was met.

The principles-based approach is more involved and takes more time and skill to apply. Output checkers must take a number of factors into account (context of research, e.g. industrial sector, occupation etc.) as well as consider what is in the public domain already that could be combined to lead to statistical disclosure. The justification to release on statistical grounds may sometimes be stronger than simply 'the results met the rule'. But possibly, this approach leads to better research outcomes as potentially more statistical outputs can safely be released.

In this Handbook we haven't chosen a preference for either approach, but instead have provided advice about making a risk assessment. Potentially this is more in keeping with the principles-based approach, as it helps output checkers make a considered judgement, using a number of factors, to consider whether to release a statistical output or not. We could have, in contrast, simply stipulated rules to follow, but we believe it is in the interest of promoting robust research using confidential sources of data not to do this.

FURTHER READING

1. Ritchie and Elliot (2015) "Principles-versus rules-based output Statistical Disclosure Control in remote access environments", IASSIST Quarterly v39 pp5-13

2. Desai, T., Ritchie, F., Welpton, R. "Five Safes: designing data access for research", University of West of England Working Paper Series, http://eprints.uwe.ac.uk/28124/1/1601.pdf

SECTION B

How to assess statistical results

Introduction: SDC

IN THIS SECTION YOU WILL LEARN ABOUT:

1. How to undertake a Statistical Disclosure Control risk assessment for a range of statistical outputs
2. What to ask analysts for when they request an output to be released

This section presents common types of statistical results. For each, a numeric example is described, with advice about how to assess statistical disclosure risk. A small box summarises the minimum information required to assess the output. Some additional considerations on secondary disclosure may follow based on the authors' experience.

It is impossible to cover every type of statistical result that an analyst could produce: Statistical Disclosure Control should not be limited to the selection that follows.

Using the guide

The following pages are each dedicated to a particular statistical output. An explanation of the statistical output is provided.

This is followed by a description of the SDC issues to consider, a ‘rule of thumb’ for checking the output, and techniques which could be applied to protect the confidentiality of data used to create the statistical output. For statistical outputs where it may not be necessary or possible to alter the results, we have not included a section on ‘Reducing Disclosure Risk’. Instead, we recommend following the guidance contained within the section – for example, ensuring the statistics are based on a sufficient number of observations.

These guidelines are not prescriptive, and should not be followed blindly by output checkers. In some cases, it might be perfectly acceptable to release statistical outputs where the ‘rule of thumb’ is not met, because an output checker reaches the conclusion that no risk of disclosure is likely. Instead, these guidelines are designed to encourage thinking about the likely risk, and to raise questions that an output checker might ask of an analyst.

We have also tried to implement the following traffic-light style approach:

🟥 Do not release unless absolutely sure it’s safe to do so

🟧 Might be safe to release, if a few sdc techniques are applied

🟩 Generally ok to release without much assessment

Descriptive statistics

MINIMUM REQUIREMENT:

- Cohort specification
- Contextual information (e.g. business data, sector, firms, time period)

Table 3:
Table of frequencies

ETHNIC BACKGROUND/AGE	16	17	18	19	20
African_Asian	0	0	0	0	0
Bangladeshi	3	4	6	5	4
Black African	3	5	7	6	8
Caribbean_West Indian	0	4	2	4	3
Chinese	0	2	1	1	0
Far Eastern	1	0	1	2	0
Indian	5	9	4	4	4
Middle Eastern	1	1	0	0	1
Mixed Caribbean_West Indian	0	0	1	0	2
Mixed Indian	0	0	0	0	0
North African	1	0	1	0	1
Pakistani	6	10	5	4	3
Sri Lankan	0	0	2	0	0
Turkish	1	1	1	0	1
White	54	135	141	146	130

Consider grouping columns or rows?

Understandable labels

Small cell frequencies

Descriptive statistics can be presented in different ways (e.g. tables, histograms, pie charts, bar charts). They typically include a measure for the centre of a distribution (mean, median or modus), a measure of the shape of the distribution (variance, standard deviation, skewness, kurtosis), and a measure for how often a certain combination of attributes is observed (frequencies, relative frequencies and counts).

Table 2:
other descriptive statistics

	MIN	1ST QUARTILE	MEDIAN	MEAN	3RD QUARTILE	MAX
Income	0	2894	6046	7178	10350	41113
Age	16	30	41	43	54	111
Turnover	140	3345993	7100730	8327156	12034046	38333022

Consider rounding or average values for e.g. 10 observations?

Precise information, probably about one observation

SDC CONSIDERATIONS

Consider the extent of detail in the output. For example, if a table simply provides a breakdown of age or gender (one-way table), or the output cross-tabulates a range of variables – for instance age and gender broken down by income level and geography (four-way table) – then the level of detail increases and attribution may occur (whereby an individual is associated with values (confidential data) and may be identified).

RULE OF THUMB

All frequencies, whether presented as observed counts or relative frequencies, should meet or exceed the threshold value 'N'. Similarly, the mean, median, mode, standard deviation or variance of a distribution/ variable should be reported based on at least N observations.

REDUCING DISCLOSURE RISK

Aggregation is often the simplest way to mitigate the risk associated with an output, other techniques include:

- **Banding/combining** columns or rows;
- **Averaging values** e.g. for 10 observations in a bracket;
- **Rounding values**;
- **Suppressing cells** (be mindful that usually two cells in a row/ column should be suppressed, otherwise the original values could be deduced).

Percentiles

MINIMUM REQUIREMENT:

- Frequency for each percentile

A percentile indicates the value at or below which a percentage of observations fall. The 50th percentile represents the median (centre) of the distribution. Analysts may also present deciles (which divide the distribution into ten equal parts), quintiles (five equal parts) and/or quartiles (four equal parts).

Table 3 shows that 95 per cent of men living in the UK had an annual post-tax weekly income of £17,734 or less, and the top five per cent had an income greater than £17,734 (time period unspecified).

SDC CONSIDERATIONS

The results are likely to be disclosive because the income values for each percentile are detailed (i.e. have not been rounded), and hence are likely to represent the income for individual men. Depending on the sample population, it could be the case that, for example, the one percentile group could actually refer to only one person, or the average for a small number of individuals in the percentile group in which, combined with other characteristics, one observation can be singled out.

Depending on the context of the analysis (e.g. population, sampling method) data subjects at the lower, upper or middle (i.e. median) of the distribution may be more easily identifiable and their income data attributable to them, or individuals may identify their own income data from the table.

RULE OF THUMB

No cell should contain less than ‘N’ observations. Check that the number of observations underlying each income value meets the threshold rule of thumb.

REDUCING DISCLOSURE RISK

Round the values (as in Table 4, e.g. to the nearest hundred or thousand). This applies to all the percentile values, including the median, and is consistent with the approach recommended for min and max values.

If the analyst wishes to show the income distribution, other options include:

- aggregating / grouping the income values into categories (e.g. income less than £12,000) to ensure the threshold rule of thumb is met. Results can then be displayed in tabular or graph form (keeping in mind that similarly produced outputs would need the same rounding technique to be applied, to avoid differencing);
- presenting the inter-quartile range.

Clear labels

Table 3:
Percentiles (unperturbed)

Percentile	Post-Tax Weekly Income (%)
1	99
5	554
10	1129
20	2346
30	3464
40	453
50	6046
60	7555
70	9287
80	11473
90	14750
95	17734
99	22997

Table 4:
Percentiles (post-SDC)

Percentile	Post-Tax Weekly Income (%)	N
1	100	10
5	600	15
10	1100	17
20	2300	11
30	3500	10
40	5000	10
50	6000	15
60	7600	19
70	9300	12
80	11500	10
90	15000	11
95	17800	16
99	23000	12

No frequencies

Median: could refer to a single observation so worth checking

Rounded figures, less likely to be attributable to specific individuals

Frequencies provided

Histograms

MINIMUM REQUIREMENT:

- Table of underlying frequencies
- Labels for axis & variables + title
- Details of data subjects and total counts

A histogram displays the frequency distribution of a variable. The width of the bars either represent class intervals or a single value, and the height represents frequency density. A density plot shows the distribution of the data over a continuous or discrete variable and it is often used to display the shape of the distribution of a specific variable.

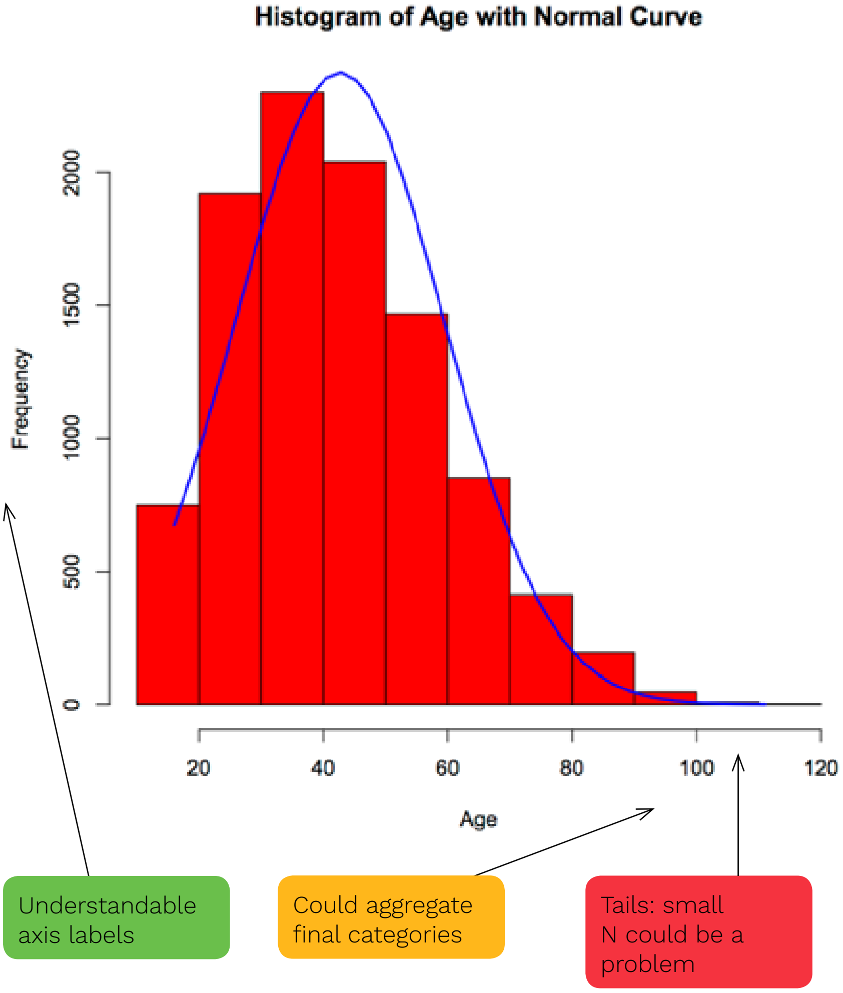
SDC CONSIDERATIONS

Graphs are subject to the same SDC methodology as statistical tables. Ideally, the easiest way to check a graph is to apply SDC to the table(s) underlying it.

For histograms and density plots, common SDC issues arise from low cell counts (below 'N' data subjects) and min/max values. As histograms and density plots are used to show the distribution of a value, low counts are often an issue, especially on the tails. Min and max are often masked in the scale of the X-axis as the start and ending points. Many statistical software programs (e.g. Stata) use by default the max value as the ending value for the axis.

Graphs should be released as fixed images (e.g. PNG), as some statistical software (e.g. Excel, Stata) can store data behind a graph. If an analyst needs to recreate the graph in a particular layout or format they should use underlying frequencies once they are checked for disclosure.

Figure 1:
Histogram



Understandable axis labels

Could aggregate final categories

Tails: small N could be a problem

RULE OF THUMB

Same rules apply as for frequency tables and min and max values.

REDUCING DISCLOSURE RISK

It is important that analysts specify the purpose of the graph, as the best options depend on the meaning behind the output. If both chart and frequency table are released, the same mitigating actions should apply to both.

If the graph is intended to show that the distribution has a long tail (i.e. there are many outliers) then analysts should cap all these values in one class. This approach can mask the maximum or minimum values. This is not necessary in cases where the minimum or maximum are a structural value or they are defined by the analyst. In the histogram presented, the minimum is a structural zero as there cannot be a lower value for age and therefore no need to conceal it. The same would apply if the data subjects were selected within an age band (i.e. only individuals less than 65 years old), where the upper limit would coincide with the max value of 65.

If the low counts are distributed outside the tails, it may be a good solution to band the variable on the X-axis into a number of classes, where each class would have enough data subjects to avoid re-identification.

If the aim is to show the shape of a probability distribution, it is possible to keep the full plot by omitting all values on the X-axis. With this solution, it is possible to relax the rules of thumb for low counts and min and max, as it would not be possible to associate any value to a specific bar or point in the graph.

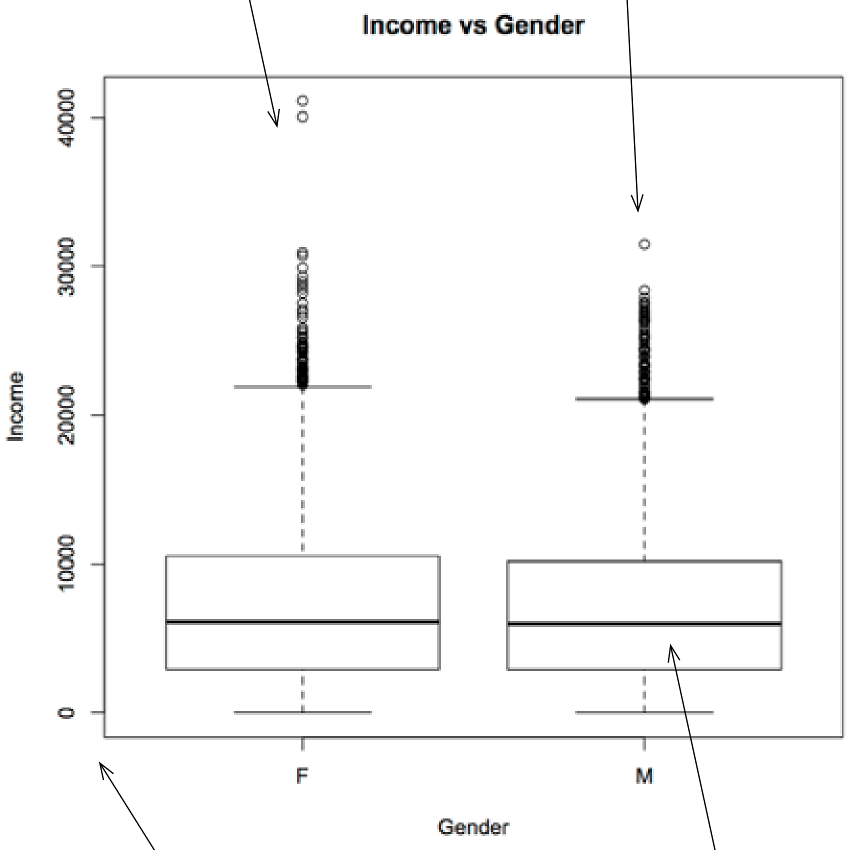
The above said, note that the scale can be reverse-engineered if even a couple of summary statistics (e.g. median and interquartile range) are released elsewhere. So it is important to check for potential secondary disclosure elsewhere.

Box plots

MINIMUM REQUIREMENT:

- Info on data source and cohort
- Labels for axis and title
- Brief description of box plot

Figure 2:
Box plot



Perhaps these observations could be grouped together and averaged out, or the range published

Outliers may be attributable to one data subject

Understandable axis labels

Information on sample size, region, occupation, etc. required

The median should be ok if based on sufficient number of observations

A box plot (sometimes known as a ‘box and whisker’) is a graphic representation of how data are distributed. Typically, box plots have:

- ‘Tails’ at either end of the plot, representing the maximum and minimum values of the distribution;
- The interquartile range (from 75th to 25th percentile), represented by the ‘box’;
- The median value.

The box plots in Figure 2 show the distribution of income of men and women from the same data source. The minimum and maximum hourly wage rates are represented by the tails on each plot; with the interquartile range of ranges represented by the boxes. The plot shows that the distribution of wages for men is ‘wider’, i.e. the minimum wage rate is lower, but the maximum wage rate is higher, for women than for men.

SDC CONSIDERATIONS

The tails are minimum and maximum values. If these values relate to single observations, then releasing these box plots may disclose confidential information about individuals.

RULE OF THUMB

Release if the analyst can demonstrate that the values of the tails are not attributable to single individual observations from the data; otherwise, recommend approaches to perturb the values of the tails.

REDUCING DISCLOSURE RISK

The minima and maxima values could be grouped or averaged. For example, instead of displaying a tail as a minimum value of hourly earnings of £7.40, band into £7 – 8 per hour (providing this met the threshold number of observations, e.g. 10).

Correlation coefficients

MINIMUM REQUIREMENT:

- Variable labels
- Number of observations

A correlation coefficient indicates the strength and direction of the linear relationship between two variables.

Correlation coefficients range from -1 to 1, with -1 indicating a perfect negative linear relationship, 0 no linear relationship, and 1 a perfect positive linear relationship.

Correlation coefficients are normally displayed in tables (e.g. Table 5).

SDC CONSIDERATIONS

Generally correlation coefficients are considered safe providing that the threshold of 'N' observations is met.

The risk may increase in cases where the correlation coefficient is exactly 1 (-/+) and descriptive statistics such as the median, percentiles, minimum or maximum are also presented in the output. For example, if there is a perfect positive correlation (=1) between turnover and employment costs for a sample of firms, and the median value for each of the variables is also presented, then those median values will relate to the one firm, since the relationship between the two variables is perfectly correlated. This may increase the risk of the firm being identified and confidential information associated with them.

RULE OF THUMB

A correlation coefficient should be derived based on at least 'N' observations.

REDUCING DISCLOSURE RISK

Correlation coefficients are rarely equal to +/- 1, as variables tend not to be perfectly correlated. Where the correlation coefficient is equal to +/- 1, and summary statistics are presented, consider suppressing or changing the correlation coefficient (e.g. =>0.80).

TABLE 5:

Correlation coefficients

Clear labels

	TURNOVER	EMPLOYMENT COSTS	INCREASE IN TOTAL STOCKS	IMPORT OF GOODS
Turnover	1.00	0.54	0.33	0.03
Employment costs	0.54	1.00	0.31	0.09
Increase in total stocks	0.33	0.31	1.00	0.19
Import of goods	0.03	0.09	0.19	1.00

Number of observations underlying each coefficient = 10,789

Number of observations provided

Factor analysis

MINIMUM REQUIREMENT:

- Variable labels
- Number of observations
- Brief description of analysis

Factor analysis is a technique for identifying factors that explain the interrelationships among data items. It is undertaken by creating a new set of summary data items: these are based on multiple data items sourced from the original dataset. The number of new data items is fewer than the original number of data items.

Factor analysis is based upon the assumption that there is a number of factors that account for the correlations amongst the original data items. If the factors are held constant, the partial correlations amongst the observed data items all become zero, and, therefore, the factors account for the values of the observed data items.

Suppose that we have a dataset of school pupil performance. We could undertake factor analysis of the dataset, by looking at the correlations among observed data items and by creating two factors:

TABLE 6:
Table of factor loadings

VARIABLE	FACTOR 1	FACTOR 2
Arithmetic score	0.40	0.89
Algebra score	0.56	0.81
Logic score	0.46	0.62
Vocabulary score	0.44	0.09
Reading score	0.50	0.12

These values are not correlations, they are factor loadings.

Comery and Lee (1992) suggest that factor loadings of greater than 0.71 indicate that over 50 per cent of the variance is explained by the factor. We could interpret from these results that Factor 2 relates to mathematical skills.

SDC CONSIDERATIONS

Owing to the methodology involved in undertaking a factor analysis, it is unlikely that disclosure of data or identification of individuals will be problematic. Consider that in the example, Factor 2, relating to mathematical prowess, could relate to one pupil in the school. This could not happen unless there was a very small number of observations in the data, because otherwise a correlation could not be established. No serious factor analysis could be undertaken without a large sample, certainly above any threshold set by a data supplier.

RULE OF THUMB

The DwB guidelines suggest that as a rule of thumb, a factor should be made up of at least two variables. This would prevent any direct correlation between a factor and an individual, unlikely as this would be with sufficient numbers of observations in the data.

Indices

MINIMUM REQUIREMENT:

- Number of observations
- Description of construction
- Interpretation

Indices provide useful aggregate statistics to describe the data, and vary hugely in their construction. Economists may produce a price index, to describe fluctuations in the price of goods or services over time. A social researcher or epidemiologist might produce an index measuring equality of access to health services across the country.

SDC CONSIDERATIONS

A ‘range’ could be considered to be an index, but this is simply a deduction of the minimum value of a data item from the maximum value. If based on observations from a small number of data subjects, the risk of disclosure will be higher than if many observations were used.

Range = Max – Min

The guidance for assessing a statistic for disclosure will be the same as for assessing frequency tables, i.e. values ought to be based on sufficient number of observations unless it can be reasonably demonstrated that disclosure of confidential information and/or risk of identification is unlikely.

By contrast, a more complex index could be created. For example, taken from the ESSNet 2010 guidelines, we reproduce a Fisher Price Index:

FIGURE 3:
Fisher Price Index

$$PF = \sqrt{PL \cdot Pp} = \sqrt{\frac{\sum_{j=1}^{m} p1,j \cdot q0,j}{\sum_{j=1}^{m} p0,j \cdot q0,j} \cdot \frac{\sum_{j=1}^{m} p1,j \cdot q1,j}{\sum_{j=1}^{m} p0,j \cdot q1,j}}$$

As you can see, this is a much more complex index in its construction, and the data used to produce the statistic will have been transformed a number of times. Owing to this transformation, statistical disclosure is unlikely to be problematic.

RULE OF THUMB

When making an SDC assessment, the output checker should consider the extent of complexity. Simple indices are more likely to be problematic than complex types.

Information about how the index has been produced is essential for this assessment.

Scatter plots

MINIMUM REQUIREMENT:

- Labels for axis & variables
- Title and cohort description

A typically displays the values for two variables and each point on the graph represents one data subject, as in Figure 4.

SDC CONSIDERATIONS

Generally, scatter plots are considered disclosive. In the example above, each point represents one data subject, which means the earnings and age of each sample member can be read easily from the graph. Presenting individual level data makes it relatively easy to identify a data subject and attribute data to them. For example, Figure 4 isolates an individual aged 35 and earning about £42,000.

The variables displayed here (earnings and age) are in their original form, as collected during the survey. They have not been transformed in any way and thus can be attributed to data subjects relatively easily. Transforming data can make it safer.

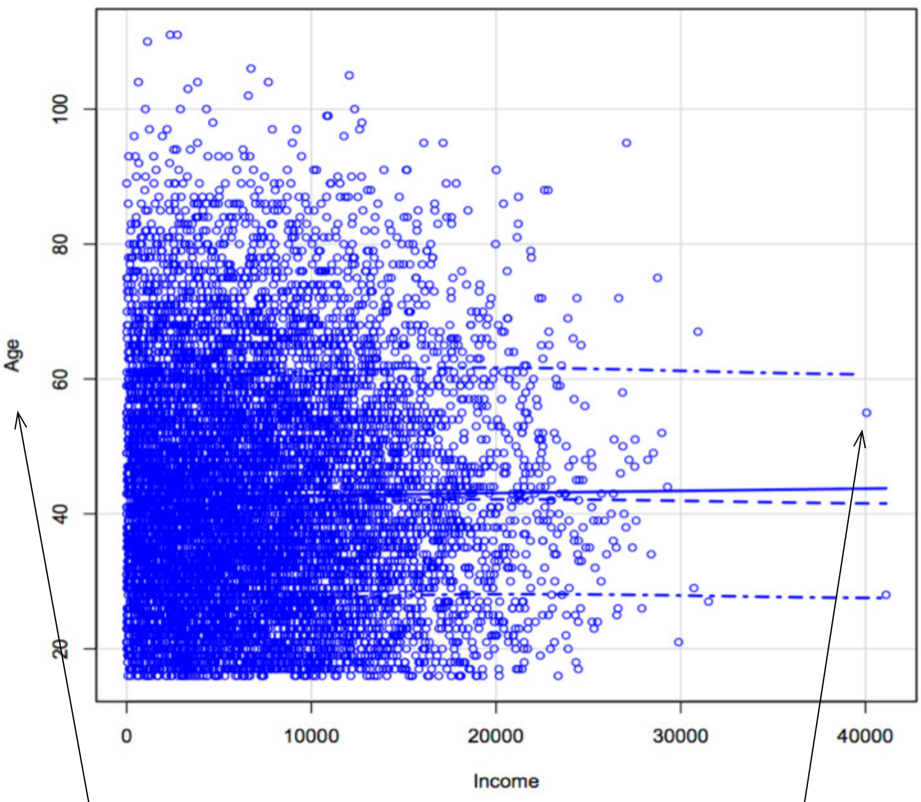
RULE OF THUMB

The rule of thumb states that no cell should contain less than N observations. In this example, it is not met as each point on the graph represents one data subject.

REDUCING DISCLOSURE RISK

One option would be to group data subjects together, to ensure that the statistics presented are based on a sufficient number of observations. In Table 7 the average earnings per hour is shown for five age groups, with the number of observations within each age group meeting the threshold rule of thumb.

FIGURE 4:
Scatter plot



Labels are clear

Individual points reported

TABLE 7:
Data perturbation for new scatter plot

Could perturb the original data and create a scatter plot

AGE GROUP	EARNINGS PER HOUR (£)	N
16-24	6.10	24
25-34	8.20	20
35-44	8.90	18
45-55	8.80	22
55 or over	10.20	14
Total		98

Symmetry plots

MINIMUM REQUIREMENT:

- Counts & clear explanation of what each point represents

A symmetry plot is a graphical technique for assessing whether a variable is symmetrically distributed. The values for the sample variable are ordered from smallest to largest. The plot then graphs the distance between the largest value and the median against the distance between the smallest value and the median. This is repeated for the second largest value and the second smallest value, then the third largest value and third smallest value, and so on, until all pairs are plotted. If the variable is symmetrically distributed all points would lie along the reference line (defined as y = x).

In Figure 5, income data for a sample of 50 women are plotted. The highest earner has an income £15,780 greater than the median, versus the lowest earner whose income is £10,230 lower than the median.

SDC CONSIDERATIONS

In the example above, each point graphs the distance from the median for two data subjects. If the median income value is known then the actual income values can be calculated for the data subjects from each group of pairs.

If there are outliers in the sample, it may be relatively easy to identify a data subject and attribute data to them (if the median is known). The income variable being displayed in the example here is in its original form, as collected by the survey. It has not been transformed in any way and as a result each value relates to a data subject.

RULE OF THUMB

No cell should contain less than 'N' observations. In this example, this rule of thumb is not met as each point on the graph represents two data subjects and the values for each data subject can be calculated (if the median is known).

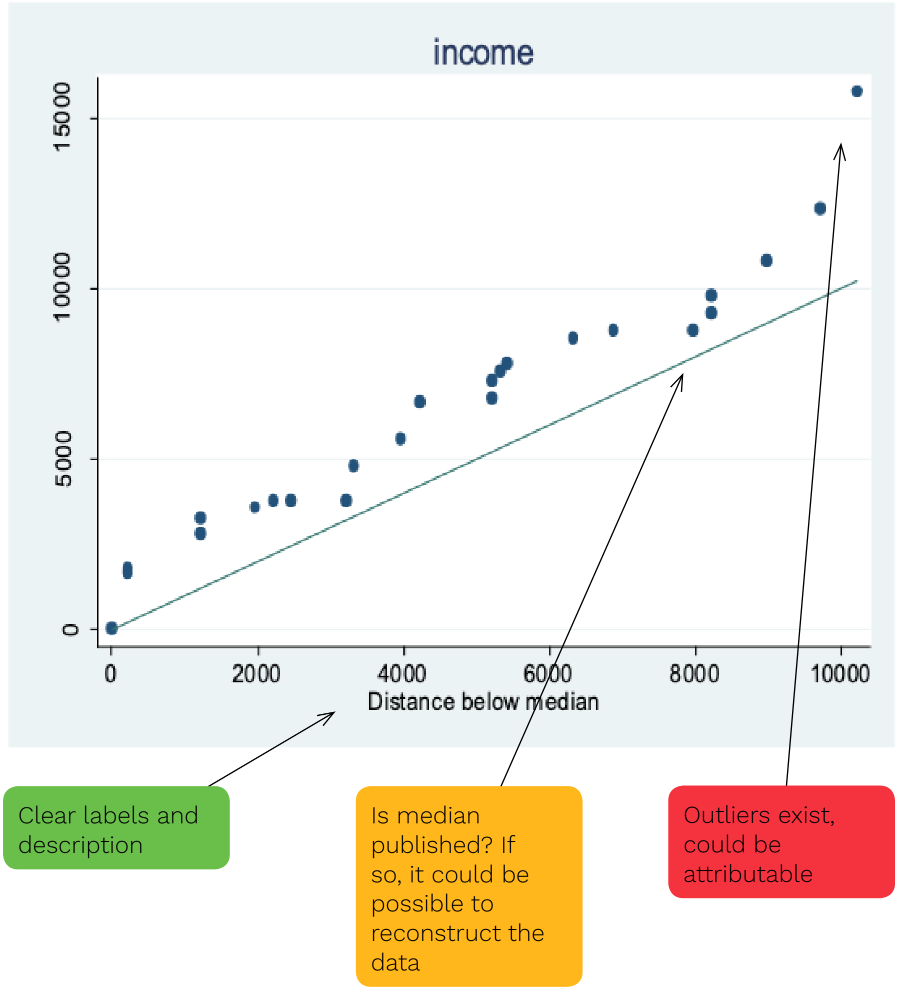
REDUCING DISCLOSURE RISK

Values could be rounded or noise added. Stata provides a ‘jitter’ function which adds random noise to data before plotting, with the option to specify the size of the noise as a percentage of the graphical area. Figure 6 shows symmetry plots where the income data has been rounded and where random noise has been added.

Standardisation is a method which transforms the values so that they have a mean of zero and a standard deviation of one. Where a standardised variable is presented in a symmetry plot, each observation’s value on the variable indicates its difference from the mean of the original variable in number of standard deviations. However this not a sufficient risk-mitigation option: if the mean, standard deviation and median are presented in the output, then these values can be used to reverse the transformation and calculate the actual value for a data subject.

FIGURE 5:

Symmetry plot (unperturbed)



Clear labels and description

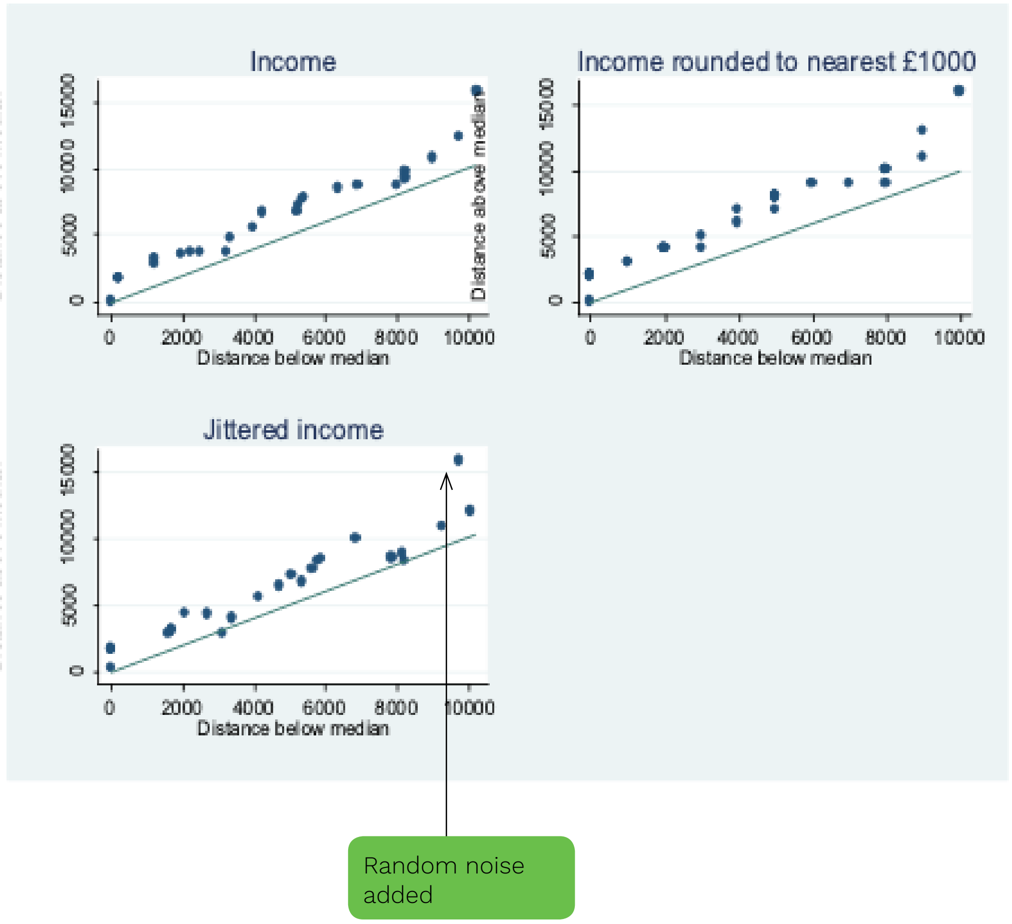
Is median published? If so, it could be possible to reconstruct the data

Outliers exist, could be attributable

If there is a data subject with an extreme value (e.g. a particularly high earner in Figure 5), then the symmetry plot may indicate them in the results. Whilst we are not able to calculate the actual income value for the data subject, the plot enables us to identify a data subject, which is problematic. An assessment of this risk therefore needs to be the symmetry plot or a similar graph (e.g. histogram) with the axis labels suppressed.

FIGURE 6:

Symmetry plot (post-SDC)



Symmetry plot of womens income. Y axis (distance above median) = Y(n=i+1) – median and X axis (distance below median) = median – Y(i) where median is the sample median, Y is the sample variable, and Y(i) indicated the ith-order statistic of Y.

Decision trees and exclusion criteria

MINIMUM REQUIREMENT:

- Number of data subject remaining for each step
- Number of data subjects dropped for each step
- Cohort specs

A decision tree diagram is like a flow diagram, in that statistical tests for each of the steps are included to assess the statistical significance of differences between the flows.

Decision tree modelling will often rely on an automated process to select the order in which variables are included in a data flow diagram. Typically, variables are chosen in such a way that the variables that explain the highest amount of variance are included in the model first. Although in some cases, the variable order can be determined a priori.

SDC CONSIDERATIONS

Can be potentially disclosive when exclusion criteria apply to small groups of data subjects, for example if very specific criteria are used, or population is small (see Figure 7 where 4 data subjects are dropped in step 3).

Potential for spontaneous identification of a data subject, or even attribution of unknown characteristics if the eligible population is well defined.

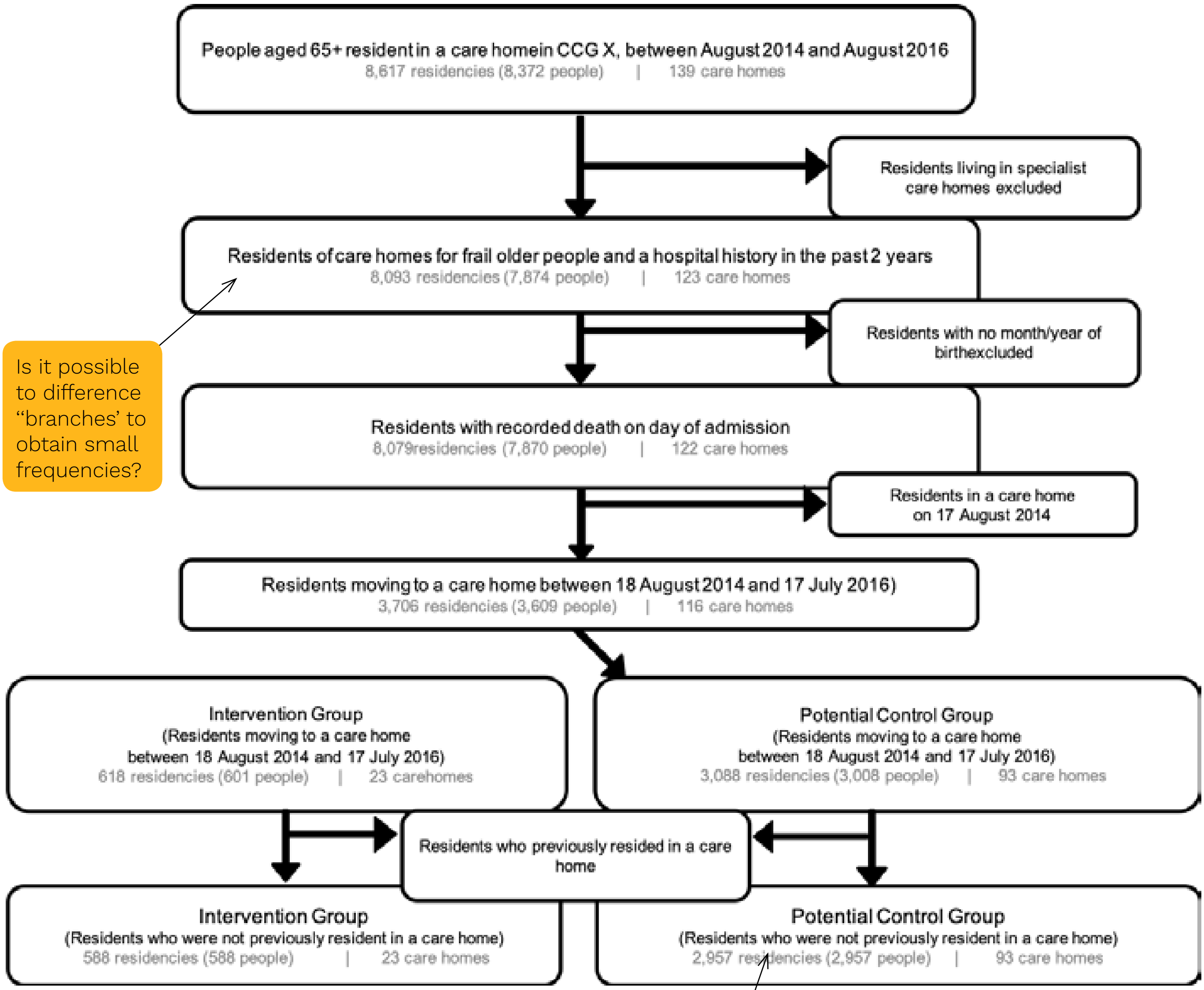
For example, from Figure 7 we already know that the data subjects are all elderly patients who have stayed in a care home in a specific area, and in step 3, we can also notice that one care home has been excluded, suggesting that all those four patients were staying in the same care home (attribution).

Secondary disclosure can arise when the same statistic is generated more than once on similar subgroups of the same population. For example, if the output in Figure 7 is replicated on a subgroup of the original population by excluding patients over 85 years old, the new output could become disclosive, even if it follows the rule of thumb for all its exclusion criteria. If the marginal differences obtained by comparing the subgroup totals for each criterion are very low, it would be possible to learn for instance that one patient who died in one of those care homes on the day of admission was more than 85 years old. Together with the other information from both outputs, this could lead to the re-identification of that patient.

RULE OF THUMB

The number of data subjects dropped in each exclusion criterion should be greater than N.

FIGURE 7:
Decision tree


People aged 65+ resident in a care homein CCG X, between August 2014 and August 2016
8,617 residencies (8,372 people) | 139 care homes
Residents living in specialist care homes excluded
Residents of care homes for frail older people and a hospital history in the past 2 years
8,093 residencies (7,874 people) | 123 care homes
Residents with no month/year of birthexcluded
Residents with recorded death on day of admission
8,079residencies (7,870 people) | 122 care homes
Residents in a care home on 17 August 2014
Residents moving to a care home between 18 August 2014 and 17 July 2016)
3,706 residencies (3,609 people) | 116 care homes
Intervention Group
(Residents moving to a care home between 18 August 2014 and 17 July 2016)
618 residencies (601 people) | 23 carehomes
Potential Control Group
(Residents moving to a care home between 18 August 2014 and 17 July 2016)
3,088 residencies (3,008 people) | 93 care homes
Residents who previously resided in a care home
Intervention Group
(Residents who were not previously resident in a care home)
588 residencies (588 people) | 23 care homes
Potential Control Group
(Residents who were not previously resident in a care home)
2,957 residencies (2,957 people) | 93 care homes

Is it possible to difference “branches’ to obtain small frequencies?

Labelled branches with frequencies explained

Survival analysis: Kaplan-Meier curve

MINIMUM REQUIREMENT:

- Frequency for each step

A Kaplan-Meier curve is a graphical representation of survival. Survival does not necessarily have to relate to mortality, but could be the time to an event (e.g. the number of days between hospital discharge and re-admission).

The graph in Figure 8 represents the number of data subjects 'surviving' (vertical axis), over time (horizontal axis). Each of the steps in the graph represents a single data subject, or group of data subjects, that do not survive at that point in time.

The Cox-proportional hazard model attempts to explain the hazard rate (inverse survival rate) using explanatory variables alongside the graph (see descriptive statistics).

SDC CONSIDERATIONS

Kaplan-Meier curve – the survival curve should be represented as a frequency table alongside the graph (see descriptive statistics).

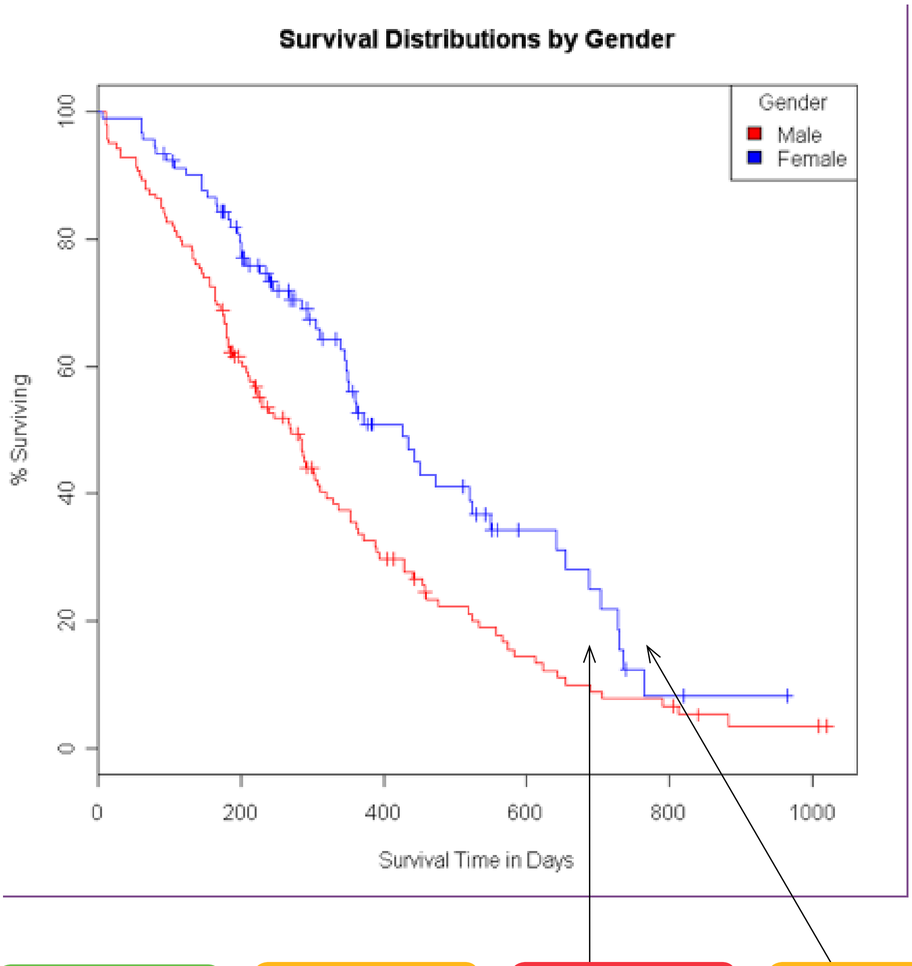
RULE OF THUMB

Each step change in the survival curve should represent at least 'N' data subjects.

REDUCING DISCLOSURE RISK

Consider banding the step changes, so that all steps in the graph represent at least 'N' observations, or consider publishing the survival curve without scales on the axes. This is especially relevant when illustrating the difference between two groups.

FIGURE 8:
Kaplan-Meier curve



Clear labels

Description and total sample size would help with assessment

Could a small number of individuals be identified via a step change?

Are there sufficient numbers of observations between step changes?

Spatial analysis (maps)

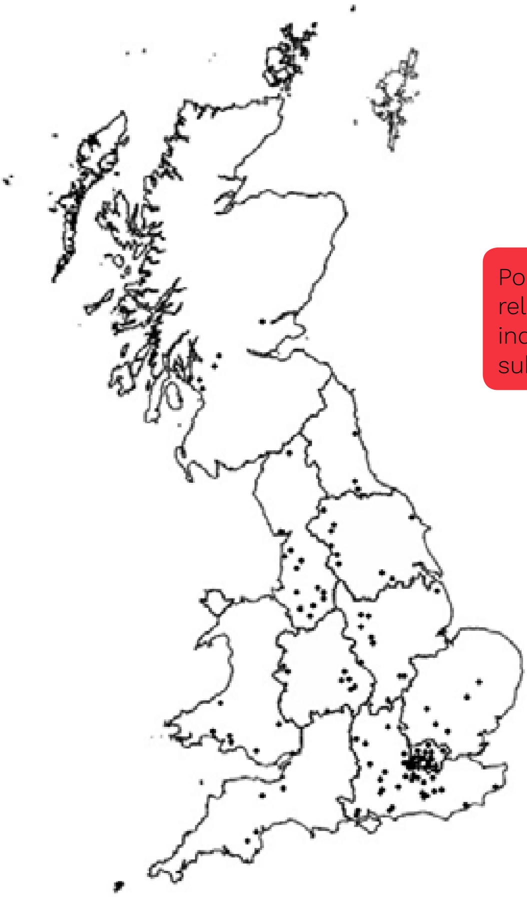
MINIMUM REQUIREMENT:

- Counts & clear explanation of what each point represents

A point map is a way of displaying data geographically. For instance, it may show the locations of businesses or types of utility.
SDC Considerations

- Each dot represents a single observation
- Are the observations ‘unusual characteristic’ (e.g. rare condition X)? Then, each point located the data subject with this characteristic.
- Dots may represent precise locations.
- Could be coupled with other informat condition, age group, to increase potent

FIGURE 9:
Point map



Points could relate to individual data subjects

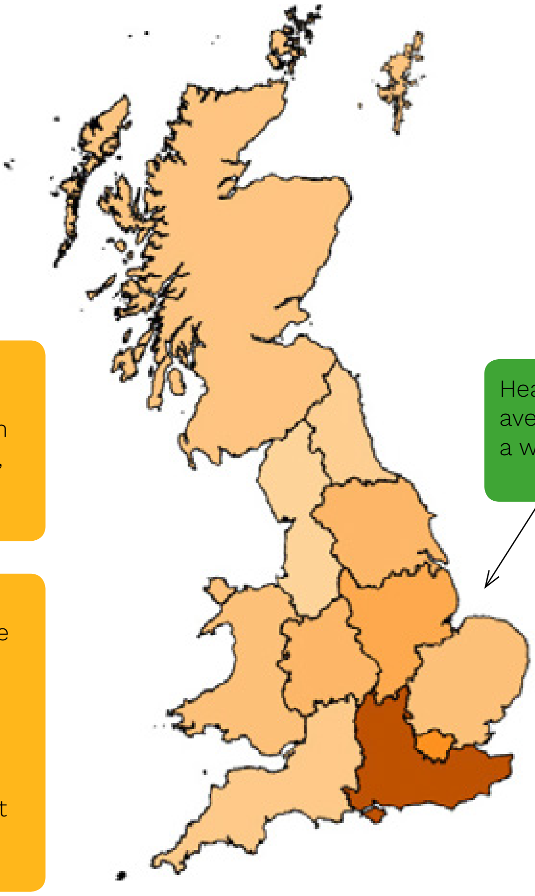
RULE OF THUMB

No cell should contain less than ‘N’ observations. Here each dot represents a single observation (N=1). If this was presented in tabular form it would very obviously not meet the threshold ‘N’ observations.

REDUCING DISCLOSURE RISK

A heat map uses colours to indicate levels of activity, intensity, concentration etc. For example, darker colours may be used to indicate high activity whilst lighter colours may indicate lower activity. Converting Figure 9 into a heat map (Figure 10) reduces the specificity of location and removes the specific numbers of patients in each area, thus significantly reducing the risk.

FIGURE 10:
Heat map



Could information be combined from public sources, e.g. Google Maps?

Depending on the scale of the map, and the nature of the data subject, it could be possible to release without disclosure occurring.

Heat map averages across a wide area

Gini coefficients

MINIMUM REQUIREMENT:

- Frequencies

A Lorenz curve measures the extent of distribution of a value over a population or sample. It is often used to assess income distribution. For example, a Gini coefficient might demonstrate that 90 per cent of the wealth in a country is owned by 10 per cent of the population.

An analyst might produce a frequency table illustrating the distribution of income for two years, and then calculate a Gini coefficient for each year. The change in the coefficient between years will determine whether income inequality has changed. In Table 8 the Gini coefficient Information on shows that inequality increased between 2000 and 2010. A Lorenz curve (Figure 11) visualises the extent of inequality, where perfect equality is represented by a 45-degree line.

SDC CONSIDERATIONS

Gini coefficients are aggregate measures. An analyst is unlikely to produce a Gini coefficient unless there are sufficient observations to create a meaningful Gini coefficient. They are usually Safe Outputs.

FIGURE 11:
Lorenz curve

Would be useful to have understandable axis labels and description of the chart

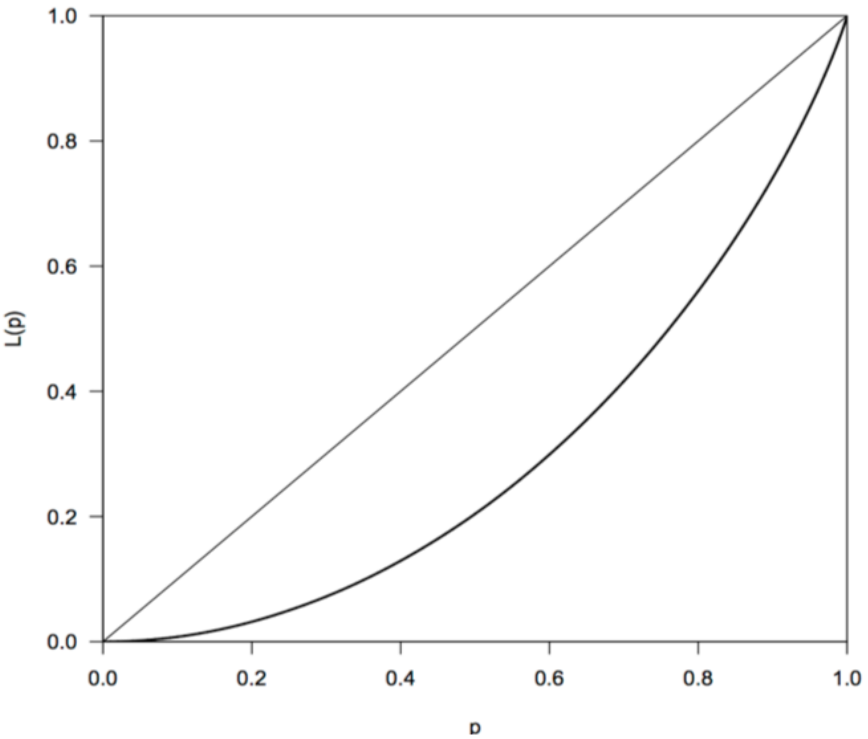


TABLE 8:

Inequality by income

INCOME BRACKETS	% HOUSEHOLDS (2000)	% HOUSEHOLDS (2010)
0 – 20,000	20	15
20,001 – 50,000	45	40
51,000 – 75,000	25	20
75,001 +	10	25
Gini coefficient	0.46	0.52

information on sample size, region, occupation, etc. required

Generally, these are aggregate statistics based on a large sample size

RULE OF THUMB

Ensure that the Gini coefficient has been calculated from at least N observations. For example, in Table 8, each income bracket will include a certain frequency of households. These should meet the minimum requirement.

For Lorenz curves, ensure that no part of the distribution refers to less than N observations

Concentration ratios

MINIMUM REQUIREMENT:

- Cohort specification
- Contextual information (e.g. business data, sector, firms, time period)

TABLE 9:
Top 3 CR for industries (unperturbed)

INDUSTRY	NO. OF FIRMS	TOP 3 CR 2010
Manufacturing	1000	0.05
Retail	9000	0.01
Digital services	150	0.2
Insurance	58	0.3
Oil refining	13	0.7

Suggests top 3 companies control 70% of the market, is there a dominance problem?

Concentration ratios (particularly, a derivation known as Herfindahl Indices), are often produced in business economics to measure how much of a 'measure' is attributable to a small number of observations. Table 9 illustrates the Top 3 Concentration Ratios for a number of industries. The 3 largest companies by market share (turnover) in the manufacturing sector accounted for 5 per cent in 2010. In the oil refining sector, the three largest companies by market share (turnover) accounted for 70 per cent of the market. In the latter case, it is likely that the industry is 'dominated' by a single observation.

SDC CONSIDERATIONS

If one observation is dominant: i.e. it accounts for the majority of the measure, e.g. turnover, income etc., then the data can be attributed to that single observation. Likely to be observed in uncompetitive industries.

RULE OF THUMB

Ensure that the ‘dominance rule’ is observed (no single observation ‘dominates’ the measure by X percent). Often, this is 40 per cent.

REDUCING DISCLOSURE RISK

Result could be perturbed: noise added, or rounded, or submitted in a ‘range’; or a larger concentration measure could be used (e.g. top 10 rather than top 3). It will depend on the effects of the change on the overall result.

Where a number of industries are compared, they could be ‘ranked’ by competitiveness rather than publishing the actual CR figures.

In Table 10, instead of producing a Concentration Ratio for each industry, the industries are now ranked, from ‘most competitive’ to ‘least competitive’.

TABLE 10:
Industry ranking by CR (CR not published)

INDUSTRY	NO. OF FIRMS	RANKING
Retail	1000	1
Manufacturing	9000	2
Digital services	150	3
Insurance	58	4
Oil refining	13	5

Shows this is a relatively uncompetitive market, but issue of dominance by one company avoided

The results may have to be suppressed entirely – instead the analyst offers a description about the extent of competition in the market without referring to results. This should be a last resort, however, given the requirement for evidence-based policy.

Regressions

MINIMUM REQUIREMENT:

- Number of observations
- Degrees of freedom
- Variable labels
- Omitted parameters
- Cohort specification

A ‘regression’ refers to the set of models for estimating a statistical relationship between two or more variables.

There are many types. Mainly there are linear and non-linear regressions:

- **Linear regression models** estimate the ‘step-change’ between variables, for example the change in productivity in response to a change in investment.

- **Non-linear models** often employ a maximum-likelihood estimate to calculate a probability, for example lone mothers are X per cent less likely to enter full-time employment following the birth of their first child.

SDC CONSIDERATIONS

Quantity requested: sometimes analysts will request to have all their regression results released (as in Table 11). At other times, they will just request a number of highlighted results which are relevant. For example, in labour market analysis, it is common to include parameters for occupation, education and industry in the regression modelling; but at conferences, an analyst would not refer specifically to these results, but instead would focus on the key results of interest.

Estimating regression parameters is an iterative process. It may take a number of iterations of model estimation (tinkering, adding and removing parameters). Analysts may request the release of regression results that may not necessarily all make it into a publication.

Why might it be disclosive?

- If the regression is undertaken on a single unit.
- If sequential regressions are undertaken but the number of observations in the cohort changes by a small number each time (and the differences in the results are significant) then this might be associated with the 'additional' observations included in the cohort.
- If the regression solely consists of categorical variables (these are variables that might take the values 0/1, True/False, Yes/No, etc.).

TABLE 11:
Table of regression estimates

TERM	ESTIMATE	STD.ERROR	T-STAT	P-VALUE
intercept	842.32	32.13	26.21	0.00
Female	-26.71	5.23	-5.11	0.00
Age 20-24	170.06	14.89	11.42	0.00
Age 25-29	111.23	14.64	7.50	0.17
Age 30-34	20.39	14.76	1.38	0.39
Age 35-39	-13.01	15.23	-0.85	0.06
Age 40-44	-28.19	15.05	-1.47	0.16
Age 45-49	-21.26	14.98	-1.42	0.04
Age 50-54	-30.52	15.09	-2.02	0.00
Age 55-59	-72.56	15.38	-4.72	0.00
Age 60-64	-111.01	15.77	-7.04	0.00
Age 65-69	-122.34	15.68	-7.80	0.00
Age 70+	87.95	31.77	2.77	0.00
Marital status (Married/Cohabiting/Civil Partner)	34.81	7.50	-4.64	0.00
Marital status (Divorced/Widowed/ Previous Civil Partnership)	57.83	10.30	5.61	0.00
Highest Qualification No Answer	168.35	55.01	-3.06	0.00
Highest Qualification Degree	121.51	30.04	-4.00	0.00
Highest Qualification GCE A level	143.01	31.19	-5.15	0.00

Easy to understand labels

Should not be undertaken on one observation (time series)

RULE OF THUMB

Generally, regression results can be released; a check for degrees of freedom (should be at least N), and that sequential regressions do not differ in counts of observations of less than N) should be undertaken.

REDUCING DISCLOSURE RISK

Ensure sufficient number of observations, within and between model estimations. In practice, this is very rarely an issue to be concerned about.

Residuals

MINIMUM REQUIREMENT:

- Axis labels

Residual plots are a graphical representation of residuals (or errors) following the estimation of a regression model. Residuals are often plotted against one of the covariates used in the model (as in Figure 12).

SDC CONSIDERATIONS

Individual residuals should never be reported. However, the shape or distribution of residuals (observed errors) following a regression can help describe the fit of a particular model or the heteroskedasticity in the data.

Important considerations when assessing a residual plot are outliers, and what variable is used on the x-axis.
In order to attribute a single residual to a specific data subject, one has to successfully order all observations along the reported x-axis. This is often difficult to achieve with a large number of observations and one plot. When multiple residual plots are presented following estimation of the same model this increases the disclosure risk since the residuals are re-arranged along the x-axis so outliers can be easily identified.

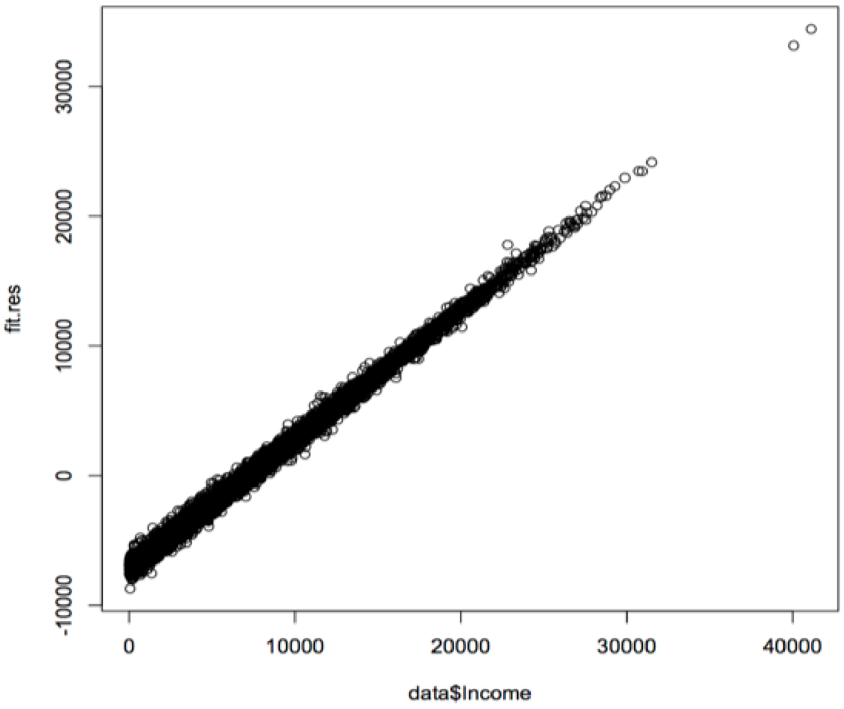
RULE OF THUMB

As each residual represents a single observation, reporting residual plots should largely be avoided.

REDUCING DISCLOSURE RISK

Consider describing the shape of a residual plot, or the conclusion drawn from inspecting it, rather than reporting the residual plot itself.
If a residual plot is needed, risk can be mitigated by removing the scale from the axes, and where possible use a covariate on the x-axis that is difficult to observe outside the dataset.

FIGURE 12:
Plot of individual residuals



Shape of distribution is ok to release

Perturbation could be a solution to enable release

Individual points

Margin plots

MINIMUM REQUIREMENT:

- Number of observations
- Degrees of freedom
- Description of dependent and independent variables

A margin plot graphs the predictive margins from a regression model. Margins are generated from regression models and show the predicted mean value of the dependent variable for the categories or values from the independent variable(s). They can be displayed in tables (e.g. Table 12) or plots (e.g. Figure 13).

SDC CONSIDERATIONS

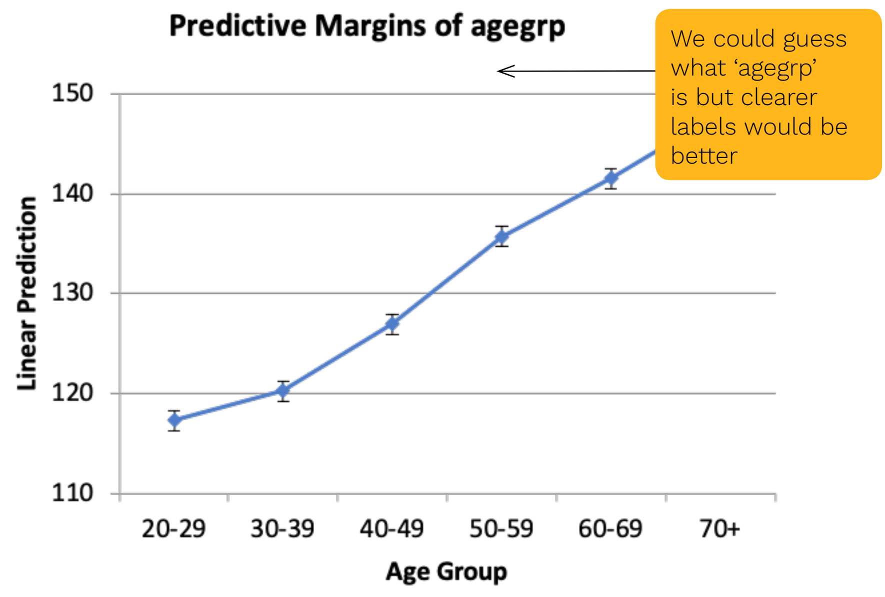
Generally considered safe as margin plot values are predictions / estimates.

RULE OF THUMB

All modelled results should have at least the threshold ‘N’ number degrees of freedom and at least ‘N’ units used to produce the model. Model should not be based on one unit (e.g. time series on one company).

FIGURE 13:

Plot of predictive margins



We could guess what ‘agegrp’ is but clearer labels would be better

TABLE 12:

Table of predictive margins

	DELTA METHOD DELTA METHOD					
Agegrp	Margin	Std.Err.	t	P>[t]	[95% Conf. Interval]	
20-29	117.2684	0.419845	279.31	0.000	116.4454	118.0914
30-39	120.2383	0.502813	239.48	0.000	119.2541	121.2225
40-49	126.9255	0.56699	223.86	0.000	125.8141	128.0369
50-59	135.682	0.562859	241.06	0.000	134.5787	136.7853
60-69	141.5285	0.37812	374.30	0.000	140.7873	142.2696
70+	148.1096	0.644507	229.80	0.000	146.8463	149.373

Is model based on sufficient numbers of observations?

Test statistics

MINIMUM REQUIREMENT:

- Number of observations

Test statistics are commonly used in statistical analysis. The nature and derivation of a test statistic will depend on what hypothesis is being tested. Some examples of test statistics include: t-test, R2, F-test, etc. Typically, a test statistic is a complex representation of a large number of observations, used to test differences between two or more groups of observations, test whether a particular parameter is different from zero or test whether an observed distribution resembles a theoretical distribution.

SDC CONSIDERATIONS

Test statistics are often calculated based on a large number of observations. The risk associated with test statistics is further mitigated by the complexity of the test statistic itself.

Note: test statistics of two marginally different samples can result in disclosure, by looking at the difference between two statistics.

RULE OF THUMB

A test statistic should be derived based on at least
'N' observations.

SECTION C

Organisation and managing requests

Introduction

This Handbook is designed both for staff working in services that provide access to confidential data, and the users of those services; and staff working in organisations that provide a secure data environment for internal use for those staff, who are likely to be assessing the results for statistical disclosure risk themselves.

Section C is primarily aimed at the former: for staff working in a service that provides secure access to confidential data, and for who the users are likely to be external (although a well-funded 'in-house' solution might well work along the same principles: for example, the ONS Secure Research Service provides access to internal ONS staff in the same way that it does to external analysts).

An important reason exists for making this distinction. The way Statistical Disclosure Control is organised can have a bearing on the safe release of results. If managed inefficiently, there is a higher risk that something will be missed, or an error made, resulting in the publication of a statistic that could reveal the identity of a data subject and/or some confidential information.

However, the way that organisations manage statistical results requests, and the process of Statistical Disclosure Control, will vary according to its own circumstances. For example, an 'in-house' solution may only have a need for Statistical Disclosure Control assessments to be undertaken once or twice a week, depending on the number of projects being worked on. This contrasts to a large service, such as the UK Data Service, which receives on average five to 10 statistical results requests from users every working day.

Whatever the circumstances, the advice and guidance which follows may be applicable to both situations. They reflect the shared best practice and wisdom of staff working in both environments over a number of years.

Implementing SDC as an organisation

IN THIS CHAPTER YOU WILL LEARN ABOUT:

1. How to encourage analysts to produce results which can be easily assessed for statistical disclosure
2. Why it is important for two people to check results
3. Why it is important for staff undertaking checks to be independent from the analysis being undertaken
4. How to manage workloads and pressure
5. The importance of keeping records and auditing
6. The wider context of undertaking Statistical Disclosure Control

Statistical Disclosure Control represents a valuable tool to any organisation that routinely produces statistical results from sensitive data. Having a robust approach to statistical result checking provides a framework to ensure that information released in the public domain does not breach data subjects' confidentiality and at the same time satisfies the risk appetite of data providers.

The design of any SDC process should reflect the organisation's aims and analytical purposes, and has to be consistent with the overall approach to data security. There is no prescribed way to implement this process as this is largely bespoke to the type of data access provided. This section provides an overview of some specific aspects to take into consideration when setting up or running an SDC service, the challenges associated, and some practical examples based on best practices adopted by a number of data centres in the UK.

INCENTIVISING GOOD RESULTS

A successful SDC process relies in large part on the quality of the results generated by analysts. Clear and easy-to-understand results enable checkers to provide better assessments and reduce the time and effort of statistical results checking.

An organisation should encourage ways to incentivise analysts to produce good results. A basic training on Statistical Disclosure Control covering the most common types of results can be delivered to all analysts accessing sensitive data. As a further measure, analysts can be asked to apply these principles to every result they produce. We recommend that guidance on what information is needed by checkers for clearing each type of statistical results is provided to analysts in advance.

FOUR-EYES PRINCIPLE

It is generally best practice to have statistical results checked by two different checkers. These can be specific members of an organisation or internal analysts but, in general, each statistical result request should be scrutinised by at least one person not involved with the original statistical result. Each check should be carried out independently of the other checker, and the decision to release a set of results should be taken jointly to ensure consistency.

Large organisations or data services may experience requests for a variety of results, some of which may prove challenging for a checker to assess. Having a second pair of eyes to review the statistical results may help with building confidence towards the decision whether to release or not, and mitigate the risk of mistakes by requesting a second opinion.

For other organisations, for example that provide 'in-house' secure data environments for their staff, this may not be necessary (if, for example, only one or two sets of statistical results are released on a rare basis). This is because more time is available to the staff in assessing the statistical results for SDC. In a busy service such as UK Data Service, the four-eyes principle is more appropriate.

INDEPENDENCE OF CHECKERS

Any consistent SDC process relies on the ability of the person in charge of checking statistical results to provide an impartial assessment. As an organisation, it is important to ensure that there is an adequate level of segregation between the analyst who produces the statistical results and the person responsible for checking it.

Small organisations and in-house services often tend to rely on their own analysts to check statistical results. In these circumstances, it is important that this role is kept separate from the work produced as an analyst. A basic rule is not to allow checkers to release their own statistical results but to require a second assessment from another analyst, possibly from another project. In small teams, this may prove challenging as all analysts may have a potential conflict of interest if they are working on the same project. In this case, a robust auditing process and random spot checks can provide a good solution to monitor the quality of the statistical results released.

MANAGING WORKLOAD AND PRESSURE

As SDC is the final safeguard before releasing potentially sensitive information in the public domain, an organisation should allocate enough resources to ensure that checkers are in a good position to make their judgment. Stress, pressure and high workloads may quickly lead to the release of disclosive results. As an organisation, it is important to provide an environment where checkers feel confident to reject bad results and do not feel pressured by analysts to make exceptions.

Setting up a Service Level Agreement (SLA) is a way to release some pressure from the system and allow checkers enough time to assess each request adequately. Analysts need to be aware that requests should be submitted in a timely way and that their work schedule should account for the time necessary for a request to be reviewed. In some cases, scheduling a specific time slot for statistical results checking can encourage analysts to concentrate all their statistical results in one request, avoiding fragmentation and facilitating the work of the checker.

A system of rotation between checkers can also improve the performance of the process, and can avoid favouritism or unbalanced workload towards more friendly checkers. This would also encourage checkers to work on a different variety of statistical results and ultimately can increase consistency in applying SDC.

Ensuring that output checkers work on other tasks during the working week will help avoid 'output fatigue' and burnout, which could lead to a reduction in service quality.

Ideally, staff should check outputs with different members of the team; if a pair of output checkers are both making a consistent mistake together, this may go unnoticed.

Some services may consider that not every set of statistical results need to be checked. For example, regression results contain a very small risk of disclosure; by contrast, more risk is often inherent in tables of frequencies and descriptive statistics. It could be that checkers only undertake SDC of these latter statistics, and automatically release, for example, regression results without question. It may be an understandable position to take: however, SDAP recommends that every set of statistics is checked, even if just a cursory glance is applied for results such as regressions. A safeguarded approach would be to encourage prolific-requesting analysts whether they require so many statistical results to be released, especially if they know that they are unlikely to use or refer to the majority of the results requested for release.

RECORD KEEPING AND ACCOUNTABILITY

As a tool to mitigate risk, it is important to keep a record of all statistical result requests in order to have the evidence to assess whether the process in place meets the organisation's needs. In particular, there are a number of reasons why keeping a comprehensive record of the SDC process may be useful for an organisation.

First, this helps monitor the volume and type of the statistical results requests while generating information on the quality of the statistical results and the decisions of checkers, with scope for continuous improvement of the process. It is also a tool for identifying who is accountable for each request and to ensure that the SDC process is followed correctly and with no confusion.

Second, a historical log of the SDC activity offers a useful knowledge base for all checkers, especially when facing results or data sources they are less familiar with. It also provides reassurance on past decisions of release in the case of audits from data providers. Finally, it is an easy way to generate logs on risk management for external or internal audits (e.g. ISO 27001, Data Security and Protection Toolkit) and for enabling spot checks. At the same time, it provides a consistent approach with the conditions required by the forthcoming data protection regulations (e.g. privacy by design).

There is no general rule on how to keep a record of the SDC process. Audit logs can be tailored to fit an organisation's needs, the volume of the service, the resources available and any security or information governance requirements (e.g. ISO 27001, IG Toolkit). It is important that the process does not become too burdensome to follow, as that may lead to misreporting from both analysts and checkers.

In general, it is good practice to keep a record:

- of all statistical results requested;
- the checker's decision (i.e. statistical results released, rejected or withdrawn);
- who requested the statistical results;
- who checked it;
- any issues or amendments applied;
- and the reasoning behind each decision;
- date when the request was made, approved, and statistical results released.

Additionally, it could be useful to keep a copy of the statistical results released in their original form. This is to allow spot checks and minimise secondary disclosure that could arise from multiple copies of the same statistical results being released.

AUDIT

A key role of the risk management of statistical results checking relies on a continuous audit of the SDC process.

It is advisable to have regular internal audits of the SDC activity, in order to monitor the quality of the statistical results released and to identify any issues that may lead to a security problem. For instance, the rejection rates of statistical results can represent a good indicator of a problem. An unusually high rate may be related to poor quality of statistical results from a particular analyst or to the need of further training for some checkers, or it may ultimately indicate that SDC is applied incorrectly. On the other side, a very low or null rejection rate may signal a failure to comply with the SDC procedures, leading to the risk of releasing potentially disclosive results from the system.

An internal process of revision allows an organisation to identify scope for continuous improvement and enables a constructive debate on the system adopted. On a similar line, external audits can provide a valuable opportunity to get a fresh opinion on the SDC process. It is important to design audits with the aim to improve the SDC process and to encourage checkers and analysts to report issues and near misses. It might be beneficial to run regular meetings for checkers to discuss particularly challenging statistical results, identify training needs and share expertise on the way they apply SDC. Spot checks on the statistical results released can be used as an additional tool to monitor the performance of the process, with statistical results deemed too disclosive being withdrawn from the public domain.

The concept of statistical disclosure is dependent on the context in which a statistical result is released, the interpretation of the checker and the risk appetite of that moment. These factors may change over time, especially when more statistical results of the same kind are released leading to an increase in the risk of secondary disclosure. Random spot checks and audits provide a way to review this risk and ensure that the approach is consistent over time and in line with the organisation’s goals.

UNDERSTANDING THE WIDER LANDSCAPE

Statistical Disclosure Control is a widely applied tool, which addresses an increasingly complex legislative framework on data confidentiality and information governance. In this fast changing landscape, an organisation can struggle to identify the right set of skills for this type of role. At the same time, checkers can feel challenged by new statistical methodologies and novel data sources, which can potentially bear new and unknown forms of disclosure.

In this context, it is important to identify training opportunities and keep checkers abreast of best practices, changes in legislation and new SDC techniques. In addition, an organisation should encourage checkers to share and discuss informally particularly challenging statistical results and the issues faced. Ultimately, Statistical Disclosure Control is a tool that reflects the approach of an organisation to risk management. For this reason, it is important that checkers feel confident when reviewing statistical results and have a route to share any issues or seek expert advice.

Managing analysts

IN THIS CHAPTER YOU WILL LEARN ABOUT:

1. The importance of taking a consistent approach to SDC
2. How organisations providing a similar service, or providing access to similar data, could work together to take a consistent approach

Receiving a large number of frequent requests for releases of statistical results from analysts can introduce challenges. For example, should requests be prioritised? How can consistency in how SDC is applied be achieved? Importantly, how can relationships with analysts be nurtured to achieve the most efficient outcome. This section considers a number of issues that affect analysts as well as staff.

CONSISTENT APPROACH TO SDC ACROSS ORGANISATION/ CHECKERS

Where there are multiple staff conducting SDC in an organisation, it is important that a consistent approach is applied by each checker. This includes:

- assessing statistics for Statistical Disclosure
- Undertaking SDC in the same way;
- ensuring that the type of statistical results submitted by the analyst meets the organisation's requirements.

Organisations will differ in their approach to assessing statistics for SDC and in relation to the type of statistical results they require analysts to submit (see 'What are good results?' below for recommendations). However, it is key that checkers apply the service's approach consistently. If this is not done then analysts may favour particular checkers (e.g. due to them being more lenient), increasing the risk of unsafe statistics being published.

Checkers should liaise with analysts in a consistent manner and ensure that messages are communicated consistently. For example, where additional information is requested from an analyst in order for the checker to make an assessment about whether a statistic is safe or not, the checker should clearly explain the reason/s for this. Likewise, if a statistical result is deemed unsafe and revisions are required, the checker should ensure that the analyst understands the reason/s for this and is able to make the required changes to produce a safe set of statistical results.

Finally, it is important that the approach adopted by the organisation to protect the confidentiality of data subjects is consistent over time. We would recommend that the above are monitored through regular audits, and where issues are identified, that changes are implemented (e.g. through staff training).

There could be advantages to be realised where services offer access to similar data types, or even provide access to the same analysts, to provide a consistent approach to SDC. It would be more efficient because analysts wouldn't have to learn 'two or more' approaches and remember to apply depending on which service they were accessing data for. Convergence could be an aim of different services.

What is a good set of statistical results?

IN THIS CHAPTER YOU WILL LEARN ABOUT:

1. What constitutes a 'good' statistical output
2. What information to ask an analyst to provide when they submit an output

Organisations will have different requirements relating to the type of statistical results they will check. Here are some suggestions about what constitutes a good set of statistical results:

- Well explained
- Description of the project
- The dataset/s used
- Sample selection criteria
- Method
- Description of the variables
- Description of the results
- Neatly presented
- Tables and figures numbered
- Variables clearly labelled
- Easy to read etc.

INCLUDES THE REQUIRED INFORMATION

This is the information that needs to be provided by the analyst so that the checker/s can make an assessment about whether the statistics are safe. This includes information such as number of observations, as well as clear descriptions of the variables. See 'Minimum Requirement' (e.g. information that needs to be provided by analyst) under each SDC technique in Section B of this Handbook.

MINIMUM AMOUNT NEEDED TO BE RELEASED

We recommend requesting that analysts think about and select the statistics they need to present their findings, rather than presenting all of the statistics they have run during their analysis.

NOT A LOG FILE

We suggest that log files or statistical results pasted from a log file are not released. This is because these types of statistical results do not meet the above requirements (e.g. well explained, neatly presented etc.) and are also likely to include statistics not required for publication.

REASON FOR RELEASE

Analysts should set out their reason for release (e.g. journal publication, presentation).

Organisations may wish to use a request form to record analysts' requests and capture some of the above information where appropriate. This could be available online or within the Safe Setting.

WHY ARE GOOD STATISTICAL RESULTS IMPORTANT?

Where good statistical results are not produced there is likely to be a higher risk of disclosure.

For example, ensuring that statistical results are well explained and neatly presented means that checkers understand the statistics and can make an assessment about whether they are safe quickly and easily. Where a bad set of statistical results is submitted (e.g. with poor description and presentation), checkers will find it difficult to understand the results and although additional information can be obtained (e.g. via requests for further information from the analyst), the statistical results will be harder to check and the risk of disclosure therefore likely to be higher.

Requesting that analysts submit the minimum amount needed for publication helps to reduce the risk of secondary disclosure. Where statistical results with a large number of statistics are submitted (e.g. numerous descriptive statistics), the risk of secondary disclosure (e.g. disclosure through differencing) will be higher.

To ensure good statistical results are produced by analysts, we recommend that organisations have a system in place to incentivise good behaviour and avoid bad (i.e. statistically risky) statistical results being produced and published. This could include prioritising the release of 'good' results over 'bad' results, and explaining to analysts of both sets of results why they were prioritised. See Ritchie & Welpton (2013) for similar approaches to this.

THE GOOD AND THE BAD OF OUTPUT CHECKING

In August 2018, SDAP organised a special workshop on SDC, bringing together a number of SDC practitioners. Each participant wrote down an example of a 'good' output request and also an example of a 'bad' output request.

The results are shown overleaf. In general, good outputs were considered to be:

- easy to understand (clear labels on graphs etc.);
- well explained (methodology and interpretation of results).

By contrast, bad outputs:

- contained little or no explanation about what the results showed;
- were often little more than log files created by analysts as part of their daily work.

The good

Box plot requested with explanation about why outliers were safe to release

Asked for advice before making request

Results included separate calculations to prove that ‘dominance rule’ had been met

Provided draft journal article so easy to understand context of results

Only requested what was required

Clear explanation of the results

Results presented clearly, ready for publication

Analysis plan accompanied the results, so it was easy to understand the results

Made request in plenty of time for the presentation

Data citation provided

Clear explanation of what the results meant

Frequency counts provided for graphs

Willingness to explain methodology

Easy to understand variable names

The bad

Researcher requested many outputs to be returned to them the same day

Researcher wanted to release a dataset to share with others

Researcher requested 800 files for release on the same day

Poor table arrangement/ formatting meant that it was really hard to check the results

Variable names in file meant nothing to me, but probably meant something to the researcher

Output consisted of results with accompanying explanation, but it was written in a different language

Output request consisted of charts and scatter plots which had no labels, or any other information

Researcher had finished his PhD but kept asking for outputs because he had to do revisions

PhD student was unsure of the methods and data, but supervisor wouldn't request access, so PhD student kept requesting small outputs

Researcher requested a huge volume of results for release, and was very pushy about the request

Output requested with no explanation

Researcher requested a large series of tables, many of which could be subtracted from each other to reveal small frequencies

Output included the postcodes for each individual in the data

Researcher requested a graph which wasn't labelled and not explained

Output requested was a huge log of the day's analysis and exploration, with no explanation about what was requested

Researcher asked for a dataset to be released (it was disguised as a 100-column table)

Graph requested was in an Excel sheet: underlying data were included

Researcher had poor knowledge of statistics software

Underlying frequencies not provided

Managing expectations

IN THIS CHAPTER YOU WILL LEARN ABOUT:

1. How to build a good relationship with analysts who produce statistics
2. Why it is important to collaborate with analysts
3. How new analysts can be supported

SERVICE LEVEL AGREEMENT (SLA)

A Service Level Agreement (SLA) setting out the SDC checking service offered by the organisation and the standards it maintains in the provision of that service, will help to manage analysts' expectations. Managing analysts' expectations in turn helps to manage checkers' workloads and the pressure on them, both of which reduce the risk of unsafe statistics being released. For example, if an organisation has no standard in terms of the time it takes to check a statistical result and an analyst submits a statistical result late in the day, requesting that it is released by the end of the day, this is likely to cause pressure on the checker.

Checking statistical results under pressure with little time is risky and mistakes may be made, increasing the risk of disclosure of confidential information. Having a standard in place that sets out the time required to check statistical results and which is manageable for the checkers, as well as meeting the needs of analysts, ensures that staff are not checking statistical results under undue pressure.

BE CLEAR ABOUT SERVICES AND RESPONSIBILITIES

SDC checking services will vary across organisations, as will analysts' responsibilities within the process. However, it is important that organisations clearly communicate what their SDC checking service includes, as well being clear about the exclusions (e.g. an organisation may provide advice and expert guidance on how to produce good and safe statistical results, but will not alter or change statistical results in any way to make them safe). Similarly, being clear about analysts' responsibilities (e.g. producing good and safe statistical results, being available to answer questions etc.) ensures that analysts are aware of what is required of them. Both of these helps to manage analysts' expectations and ensure that SDC checking is carried out in a non-pressurised environment.

MANAGING RELATIONS

Managing relations effectively with analysts is a key element of managing disclosure risk.

The establishment of this relationship begins at the training that analysts attend before accessing the Safe Setting. We recommend that the training provides analysts with an understanding of SDC and the skills required to produce good quality safe statistical results. It is also important to provide analysts with a clear view of their responsibilities and involvement in the SDC process, and of the joint working relationship between analysts and checkers.

Following the training, SDC should be a collaborative process, based on mutual understanding and respect. Checkers should be willing to assess any statistic they are presented with and, where they have limited knowledge of the statistic, work with the analyst to understand it. Checkers should also provide help and guidance to analysts in making 'unsafe' statistical results safe. Likewise, analysts should take responsibility for their statistical results, taking care to produce good quality safe statistical results, and be available to discuss them with the checkers and make changes where required. Both parties should work together to identify ways in which statistical results can be released – this work should not fall solely on the checker/s.

As Desai and Ritchie (2009, p.8) demonstrate in their paper on 'Effective Researcher Management', training and involving analysts in SDC promotes a culture of understanding data security, in which analysts feel accountable for the safety and security of data. This reduces the risk of disclosure for data subjects (see the paper for a full discussion of the benefits of this system).

SUPPORTING 'NEW' ANALYSTS

Analysts that are new to working in a Safe Setting may require additional SDC support during the early stages. If a new analyst submits an unsafe or bad set of statistical results, we would recommend that the checker remind the analyst of the organisation's approach to SDC and its type of requirements.

New analysts may also require extra support to help them make 'unsafe' statistical results safe and this should be provided. Ideally these conversations should be carried out over the telephone as it gives the analyst an opportunity to ask questions and for the checker to ensure that their points have been understood. An email afterwards confirming the conversation and what has been agreed is recommended.

Additional resources may be required to support new analysts, however this is an efficient use of resources in the long-term. Providing additional support in the early stages will ensure that analysts understand their responsibilities and have the skills to produce good quality safe statistical results – saving time and effort for checkers in the long run.

Further resources

Not many resources exist which provide guidance about how to set up an SDC process, or about managing users of the service with respect to SDC. That's because not many Safe Settings exist, and they are a relatively recent invention. However, this section contains some references which staff working at Safe Settings may find useful.

EFFECTIVE RESEARCHER MANAGEMENT

Professor Felix Ritchie (University of the West of England) and Tanvi Desai (former Assistant Director of the Administrative Data Service and Data Manager at the London School of Economics) published an article about how to effectively manage users. While this article does not focus specifically on SDC, it does encourage Safe Setting staff to think about how to manage users in a positive and proactive manner, and how to go about creating incentives.

https://pdfs.semanticscholar.org/5ad1/c7b30b8310f448cabe61386c77889048a44b.pdf

OPERATIONALISING PRINCIPLES-BASED OUTPUT SDC

Professor Felix Ritchie (University of the West of England) and Richard Welpton (The Health Foundation, formerly at UK Data Service, University of Essex) drafted a paper aiming to set out the practicalities of managing SDC in a Safe Setting.

http://www.felixritchie.co.uk/publications/Operationalising%20PBOSDC%20new%20v10a.docx

ONS STATISTICAL DISCLOSURE CONTROL RESOURCES

The ONS have produced guidance on SDC for releasing microdata, including intruder scenarios, and specific guidance for releasing health statistics (aggregate tables). Although these guidelines are not specifically tailored for research outputs, the concepts explained here are useful.

https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol

GOVERNMENT STATISTICAL SERVICE

The Government Statistical Service supports statisticians working in government departments and agencies. They have produced guidance for undertaking SDC of aggregate tables created from administrative data sources.

https://gss.civilservice.gov.uk/guidances/methodology/statisticaldisclosure-control/#tables-produced-from-administrative-sources

NHS DIGITAL GUIDELINES

Researchers who access data from NHS Digital or Public Health England are required to comply with the guidelines for publishing aggregate tables using health data, stated in this document:

https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data

ESSNET GUIDELINES

Referred to earlier, the original ESSNet guidelines produced in 2010, can be found here:

http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

They were later updated as part of the Data Without Boundaries project. They can be found as part of the Deliverables in section 11.

http://dwbproject.org
https://ec.europa.eu/eurostat/cros/content/guidelines-output-checking_en

RESEARCH PAPER ON SDC AND OUTPUTS

Professor Felix Ritchie (University of West of England) has written this technical paper about SDC for outputs, and why it is necessary to consider SDC for research outputs separately.

https://wiserd.ac.uk/sites/default/files/documents//WISERD_WDR_006.pdf

GUIDE TO SDC OF OUTPUTS

Produced for the Administrative Data Research Network, this paper by Professor Felix Ritchie and Philip Lowthian at ONS describes SDC for research outputs.

https://adrn.ac.uk/media/174254/sdc_guide_final.pdf

UK ANONYMISATION NETWORK

The UK Anonymisation Network provides advice on anonymisation and data confidentiality. The site provides a number of resources which could be of interest.

https://ukanon.net

ICO ANONYMISATION CODE OF PRACTICE

The Information Commissioners Office (ICO) published an Anonymisation Code of Practice in 2012. It will shortly be updated to take account of changes since the introduction of the General Data Protection Regulation (GDPR).

https://ico.org.uk/media/1061/anonymisation-code.pdf

THE FIVE SAFES

For an explanation of the Five Safes framework for enabling safe use of data, see: Desai, Ritchie and Welpton (2016).

https://uwe-repository.worktribe.com/output/914745

Glossary

Analyst
People who are analysing microdata for research purposes and producing statistical results. In some fields of study, they would be referred to as ‘researchers’; for clarity we use ‘analyst’ throughout this Handbook.

Attribution
Where a data subject may be identified when characteristics, seemingly anonymous individually, are fitted together to form a clearer, potentially disclosive, picture.

Confidential (sensitive) information
Refers to data for which are detailed and have been collected in confidence; may be attributable to an individual, and may have direct identifies such as names and addresses removed.

Context/information
In order to aid the smooth running of the SDC process, many data providers require statistical outputs to include accompanying material alongside tables and figures, e.g. unweighted Ns, information on sample, method etc., although the extent of this requirement may vary between data providers.

Data controller
This is a term defined in data protection legislation, and refers to the organisation or individual who determines the purpose for which personal data are processed. Often this will be a data owner, but not always.

Data owner
The organisation responsible for the data. May be a data controller under the Data Protection Act, but not always. For example, the Office for National Statistics will be the data owner for data it collects through surveys.

Data providers
The organisation supplying the data to the Safe Setting. The Safe Setting might be run as a service by the data provider.

Data subject
The unit of observation in a dataset. Usually individuals or businesses, depending on the source of the data.

Five Safes
Governance framework adopted by many Safe Settings to describe approaches for managing access to data.

Four-eyes principle

The best practice principle that, where possible, statistical outputs (a.k.a. statistical results) should be checked by two people rather than one.

Identification/re-identification

Identification of a data subject is what the application of the SDC process aims to avoid. Where analysts are using highly detailed microdata, even though these data are de-identified, there is still risk of re-identification, through the combination of variables (indirectly identifiable).

Output checkers

Those responsible for checking statistical outputs (a.k.a. statistical results) created in Safe Settings for potentially disclosive issues.

Risk appetite

The level of risk a data owner is willing to take with regard to the use of their data.

Rule of thumb

A practical approach to aspects of SDC, based on experience of what is likely to mitigate against disclosure risk in most situations.

Safe Output

A set of statistics which are deemed not likely to reveal confidential information and/or reveal the identity of an individual data subject.

Safe Settings

A technologically (and sometimes physically) secure environment in which analysts access data and undertake analysis, and the statistical outputs are returned to them subject to a Statistical Disclosure Control check by staff. Also referred to as Research Data Centre, Secure Enclave, Secure Data Environment, Trustworthy Research Environment, Safe Data Haven, Data Safe Haven. Can include ‘on-site’ access and ‘remote secure’ access.

Secondary disclosure

Where two, or more, seemingly ‘safe’ statistics (e.g. tables, graphs) presented as part of an output – or even across outputs – can produce potentially disclosive new statistics when combined.

Statistical Disclosure Control (SDC)

The process applied to statistical outputs (statistical results) to mitigate the risk of potentially disclosive results leaving the Safe Setting.

Statistical output

The results of the analysis that the analyst wishes to have released from the Safe Setting and which will undergo the SDC process. Also known as ‘statistical results’.

Threshold

The number of observations underpinning the derivation of statistic that must be met to be considered ‘safe’. Throughout this Handbook we have used a commonly used threshold of 10.

Copyright

If you would like to use our work please do so but please also acknowledge us! Safe Data Access Professionals.