



# Security and Disclosure for Statistical Information

Andrew Westlake

Survey & Statistical Computing

+44 20 8374 4723

[AJW@SaSC.co.uk](mailto:AJW@SaSC.co.uk)

[www.SaSC.co.uk](http://www.SaSC.co.uk)

# Overview

- Statistics needs standard security/privacy processes and procedures
- Additional issues relating to the intended disclosure of information
- Presentation is an overview of issues and methods
- Considerable Literature
  - » Some references at end, more on web site

# Statistical Confidentiality

- Promise to Respondent
  - » Individual information will be kept confidential
  - » May be a legal requirement
- Action to preserve confidentiality
  - » Greatest risk is in transmission and processing of the original information
  - » Security systems needed for source data
- Release of Statistical Information
  - » Fitting models and estimating parameters
  - » Reporting aggregate information
  - » Users want detail and flexibility
- Disclosure Control
  - » Can sometimes infer information about an individual from aggregated results or anonymised records
  - » Particular problem when the intruder has auxiliary information



# Confidentiality of Source Data

- General security solutions
  - » Data acquisition
  - » Management while processing
- Transfer from source to statistical organisation
  - » Health or Tax records, for example
  - » Encryption of direct identifiers
    - > Conceal identity but allow matching
  - » Preservation of partial information, eg Soundex
- Problems are not specific to Statistics

# Statistical Disclosure Control

- Objective is to reveal statistical information
  - » Reporting aggregate characteristics of system/population
  - » Inferring relationships or processes
- Must still protect individual confidentiality
- Various Situations
  - » Individual Summary Tables
  - » Anonymised Individual Records
  - » Detailed Tables
  - » Tabulation on Demand

# Release of Summary Tables

- Historically the most common situation
  - » Figures are reported to support analysis and conclusions
  - » Relatively few tables and cells
- Relatively easy to protect
  - » Common approach is to suppress figures where the number of contributions is too low, or single respondent dominates a group
  - » Explore this later in more detail

# Anonymised Individual Records

- Transformed version of Source data
- Users can conduct analysis in depth
  - » Relationships between measurements and other attributes
- Protection methods
  - » All direct identification removed
  - » Reduce detail in identifying measurements
    - > Until individuals cannot be distinguished
  - » Introduce uncertainty into the identification
    - > Randomisation of responses

# Other Tabulation Situations

- Release of Detailed Tables
  - » Not limited to support for reports, but release as much detail as possible
  - » Can have multiple tables with same dimensions but different levels of detail
- Tabulation On Demand
  - » Tabulation engine with access to underlying records
  - » Automatically check results and suppress if not acceptable

# Access Control and Trust

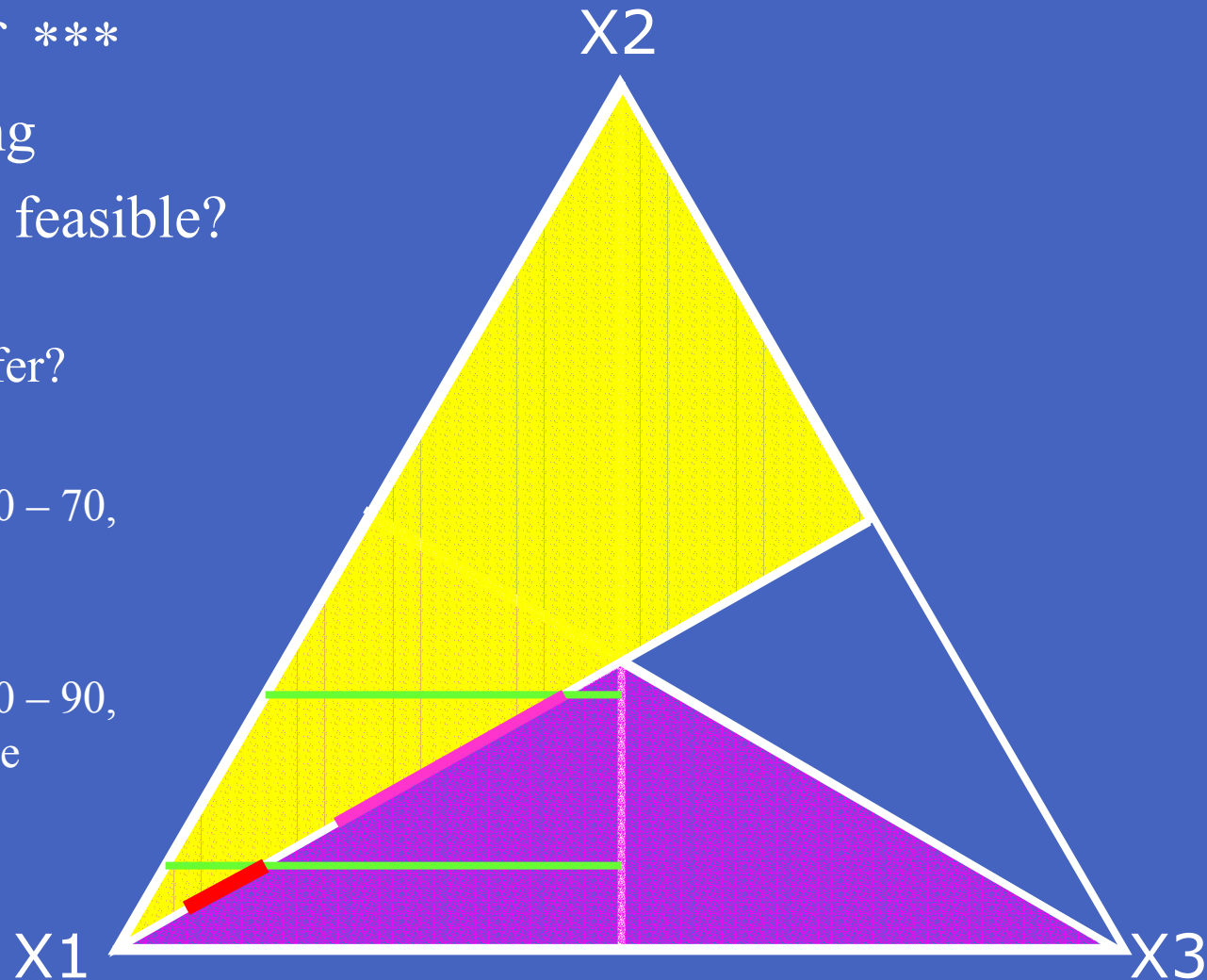
- Control the Users, not the Information
  - » Standard process
  - » Important for internal work on sensitive information
  - » Used by archives and with secondary analysis
    - > Identification of users
    - > Trust in undertakings to respect confidentiality
  - » Can be difficult to monitor that users are successfully compliant

# Summary Tables - Sensitivity

- Summary Table
  - » Dimensions are Classifications (cf OLAP)
  - » Content of cells are Measures, counts, sums, means, etc.
- What is Sensitive (or Risky)?
  - » Many proposals, no agreement
  - » N-k rule
    - > ‘Cell is Sensitive if top N respondents contribute at least k% of cell total’
    - > Widely used rule – measures Dominance
    - > Disliked by theoreticians – does not measure Sensitivity
  - » Difficult to define sensitivity (measure, not state)
    - > Objective is to quantify disclosure risk
    - > Depends on the ancillary knowledge of the intruder

# Sensitivity: Three Firms

- Consumption of \*\*\*
- All know ranking
- What values are feasible?
- What can Firm2 infer?
- If Firm2 = 30%
  - » Firm1 in range 40 – 70,  
not sensitive
- If Firm2 = 10%
  - » Firm1 in range 80 – 90,  
probably sensitive



# Summary Tables - Cell Suppression

- Cannot hide a single (primary) sensitive cell
  - » Can re-estimate the value from the margins and the other cells
- Two approaches
  - » Reduce detail in some classifications (enlarge cells)
    - > Not always feasible
  - » Suppress additional (secondary) cells
    - > At least two in every group for which an overall total (or equivalent) that includes the primary cell is available

# Secondary Suppression

- Complex optimisation problem
  - » Requires a suitable objective function
- Information Loss
  - » Various proposals
    - > Number of cells
    - > Total of suppressed values
  - » Better to use an information measure
    - > How much information is needed to reconstruct the original table
- Interesting extension
  - » Cooperating intruders who have information about different suppressed cells

# Summary Tables - Feasible Ranges

- Can estimate the range of values for each suppressed cell
  - » Simple for 2x2 case
  - » More complex for more cells and more dimensions (IP)
- Strong suggestion that the ranges should be published
  - » Since clever reader can compute them

100	5	105
3	2	5
103	7	110

103 - 98	2X7	105
0X5	0X5	5
103	7	110

# Feasible Ranges: Problems

- Bounds can be very narrow
  - » Perhaps sufficient to constitute sensitivity
  - » Should contribute to determination of secondary suppression set
- Summary cubes (OLAP cubes) with classification hierarchies
  - » Releases all possible margins
  - » More information for inference
  - » Must consider as a whole, not individually

# Anonymised Data Records

- Data records with identification removed
  - » Big demand because of scope for modelling
  - » May be a sample
  - » Identification of record gives access to all values
- Intruder knows values for some fields
  - » Not sensitive, but intruder has link to IDs
  - » Gives access to other (sensitive) values
- Identification risk
  - » Greatest in small sub-groups – all possible tables have been released
  - » Extreme values are risky because they characterise small groups

# Protecting Data Records

## Three general approaches

1. Reduce detail in classifications
  - > May not be possible
  - > Large loss of information (affects cells that are not sensitive)
2. Model-based imputation (perturbation)
  - > Fit a model (can include just noise)
  - > Replace sensitive (or all) values in risky (or all) records
  - > Dilutes (or removes) any relationship not within the model
3. Data Swapping
  - > Choose pairs of records (random, or based on risk)
  - > Swap values of a set of fields
  - > Only affects (dilutes) relationship between swapped and non-swapped fields

# Multiple Summary Tables

- Set of cubes with classification hierarchies
  - » More detail than publication tables
  - » Less Risk than anonymised records
  - » More information within selected dimension combinations
- More Combinations
  - » Can have multiple tables with same dimensions but detail in different dimensions
  - » Supports more uses
- More scope for tight feasibility bounds
  - » Must protect across the complete set of tables
  - » Harder version of previous problem

# Tabulation On-Demand

- Allow users to request tabulations
  - » Computed from data records on-demand
  - » Protect tables before return to user
- Problem with multiple requests
  - » As before, but now must protect in sequence
  - » Users may collaborate, so must protect over all requests
- Danger from differencing between requests
  - » Intrusion in the intersection

# Current Activities

- Continuing activity within Statistical Offices
  - » And related academic groups
  - » Includes some generic software development
  - » In-house development by major NSIs
- Data Mining initiatives
- CASC project supported by EU
  - » Led by Statistics Netherlands
  - » Argus software and protection techniques
- Dissemination included in AMRADS
  - » Proposed for NIPS

# Current Advice

- Do the SDC up-front on the data and tables
- Provide on-line tabulation service without disclosure control, but based on anonymised records (perhaps use  $\mu$ -Argus software)
- Also provide sets of summary tables with more detail in interesting dimensions, used through same tabulation interface
- Provide a system for users to request additional data release, which is then judged for disclosure relative to that already released

# Conclusions

- Statisticians need standard confidentiality procedures
- Separate problems for the release of statistical information
  - » Controlling risk of disclosure about individuals
- Long history of work
  - » Advances being made, but still no really clear solutions
  - » Some methods very expensive and some problems very difficult
- Some lack of focus
  - » Need clear criteria for Risk, Information Loss
  - » Too many heuristics
- How do we know whether protection methods are necessary or successful?

# References

- J. Domingo-Ferrer (Ed). Inference Control in Statistical Databases. Springer Verlag – ISBN 3-540-43614-6 (2002)
  - » Lots of useful papers
- L. Willenborg & T. de Waal. Elements of Statistical Disclosure Control. Springer Verlag (2001)
  - » A Standard Reference Work
- Project Web Sites
  - » [CASC] Computational Aspects of Statistical Confidentiality  
<http://neon.vb.cbs.nl/casc>
  - » [AMRADS] Accompanying Measure to Research and Development in Official Statistics  
<http://amrads.jrc.cec.eu.int>
  - » [NIPS] Network for the Improvement of Public Statistics  
<http://www.publicstatistics.net>

# Thank You

- Full version of paper to be loaded on web site

[www.sasc.co.uk](http://www.sasc.co.uk)

