Safe Researcher Training

Statistical disclosure control: minimising risks in output

In this section:

1. **Basic theory, using simple tables as examples**
2. **Extending this to the research environment**

Safe outputs

BASIC PRINCIPLES

What is statistical disclosure control?

- addressing residual risk in results for publication
- being precautionary, but
  - utility is important – balance with risk
  - consistent with good research
- Thought more important than advanced stats

SDC: our example dataset

Which of these variables might
- be sensitive?
- help to identify someone?

Table 1 dataset description

Variable	Description
id	random ID number
male	male y/n
age	age
white	white y/n
fragilex	has fragileX gene
diabetic	diabetes diagnosed
education	highest qualifications
abc1	socio-econonomic group
income	annual income from all sources in £
i_quartile	income quartile 1 (lowest) to 4 (highest)
imputed_value	values were imputed y/n

SDC example: small counts 1

Potential problems with this table?

	Has fragileX gene?		
	no	yes	Total
Gender			
female	85	6	91
male	58	1 x	59
Total	143	7	150

SDC example: small counts 2

Potential problems with this table?

	Has fragileX gene?		
	No	yes	Total
Diabetes diagnosed			
no	114	2 x	116
yes	29	5	34
Total	143	7	15

SDC example: small counts 3

Potential problems with this table?

	At least one value imputed		
	no	yes	Total
Gender			
female	89	2	91
male	58	1	59
Total	147	3	150

SDC example: class disclosure

Potential problems with this table?

	Income quartile (lowest 1, highest 4)				
	1	2	3	4	Total
Highest qualification					
postgrad	1 x	1 x	8	18	28
degree	2 x	6	14	17	39
college	8	18	16	3	45
school	13	9	0	0	22
none	13	3	0	0	16
Total	37	37	38	38	150

SDC example: class disclosure

- Which of these statements is disclosive?

  "all of the students aged 14+ said that they had tried cannabis at least once"

  "no nurse in the survey earns over £22.50/hour"

  "no-one in Shetland earns over £50,000/year"

  "no-one in Shetland earns over £5m/year"

  "no-one in Shetland earns over £500m/year"

- empty and full (100%) cells problematic

  irrespective of the number of observations

SDC example: structural zeros

Potential problems with this table?

Age and education of young respondents					
	16-17	18-19	20-23	24-29	Total
Highest qualification					
degree	0	0	51	64	115
college	0	25	33	57	115
school	15	18	19	41	93
none	8	7	12	17	44
Total	23	50	115	179	367

What can we do with this table?

- Suggest at least four solutions

	Income quartile (lowest 1, highest 4)				
	1	2	3	4	Total
Highest qualification					
postgrad	1	1	8	18	28
degree	2	6	14	17	39
college	8	18	16	3	45
school	13	9	0	0	22
none	13	3	0	0	16
Total	37	37	38	38	150

Option 1: hide the offending results

- Cell suppression - blanking offending cells

Highest qualification	Income quartile (lowest 1, highest 4) 1	2	3	4	Total
postgrad	<3	<3	8	18	26
degree	<3	6	14	17	37
college	8	18	16	3	45
school	13	9	<3	<3	22
none	13	3	<3	<3	16
Total	34	36	38	38	146

Calculate totals afterwards!

before

Income quartile (lowest 1, highest 4)

Highest qualification	1	2	3	4	Total
postgrad	1	1	8	18	28
degree	2	6	14	17	39
college	8	18	16	3	45
school	13	9	0	0	22
none	13	3	0	0	16
Total	37	37	38	38	150

after

Income quartile (lowest 1, highest 4)

Highest qualification	1	2	3	4	Total
postgrad			8	18	26
degree		6	14	17	37
college	8	18	16	3	45
school	13	9			22
none	13	3			16
Total	34	36	38	38	146

Option 2: change the offending results

- Rounding

	Income quartile (lowest 1, highest 4)				
	1	2	3	4	Total
Highest qualification					
postgrad	0	0	10	20	30
degree	0	5	15	15	35
college	10	20	15	5	50
school	15	10	0	0	25
none	15	5	0	0	20
Total	40	40	40	40	160

Option 2: change the offending results

- Make the data into something less disclosive

  ratios
  growth rates
  calculating proportions (but limit decimal places)
  etc

Option 3: redesign the output

- Why do we recommend this?

	Income quartile (lowest 1, highest 4)				
	1	2	3	4	Total
Highest qualification					
postgrad	1	1	8	18	28
UG/PG degree	3	7	22	35	57
college	8	18	16	3	45
school	13	9	0	0	22
none	13	3	0	0	16
Total	37	37	38	38	150

Your choice…

- You know what's important

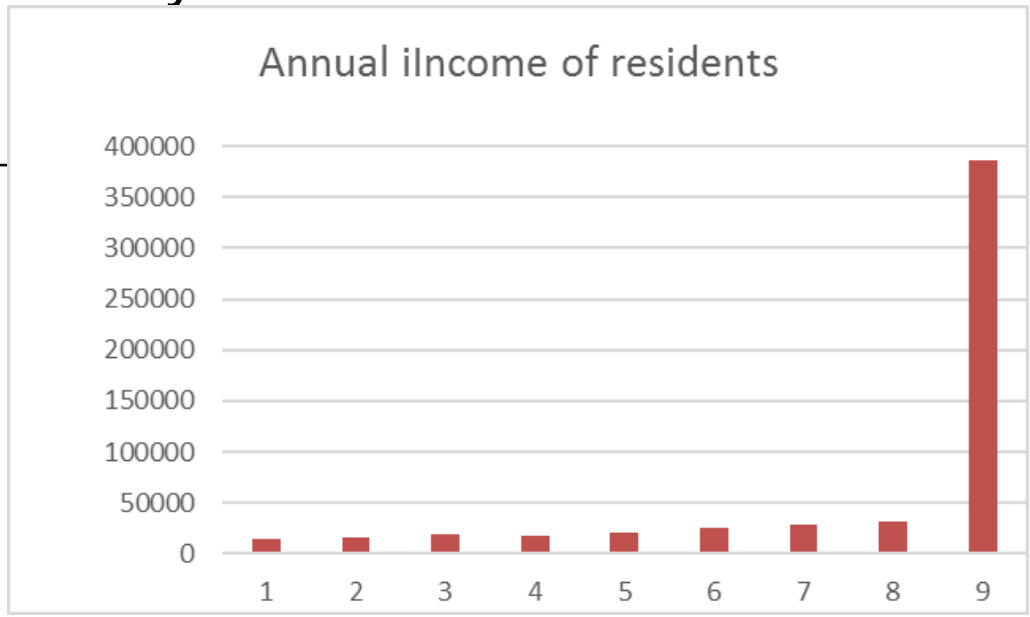
  $\Rightarrow$you decide what SDC methods to use

- User support team can help

SDC example: dominance

Potential problems with this table?

	Hobbiton		Bywater	Overall
	N	Mean income		
Degree	9	£62,384		
College	11	£42,367		
School/ none	13	£16,017		
Overall		£37,446	£41,531	

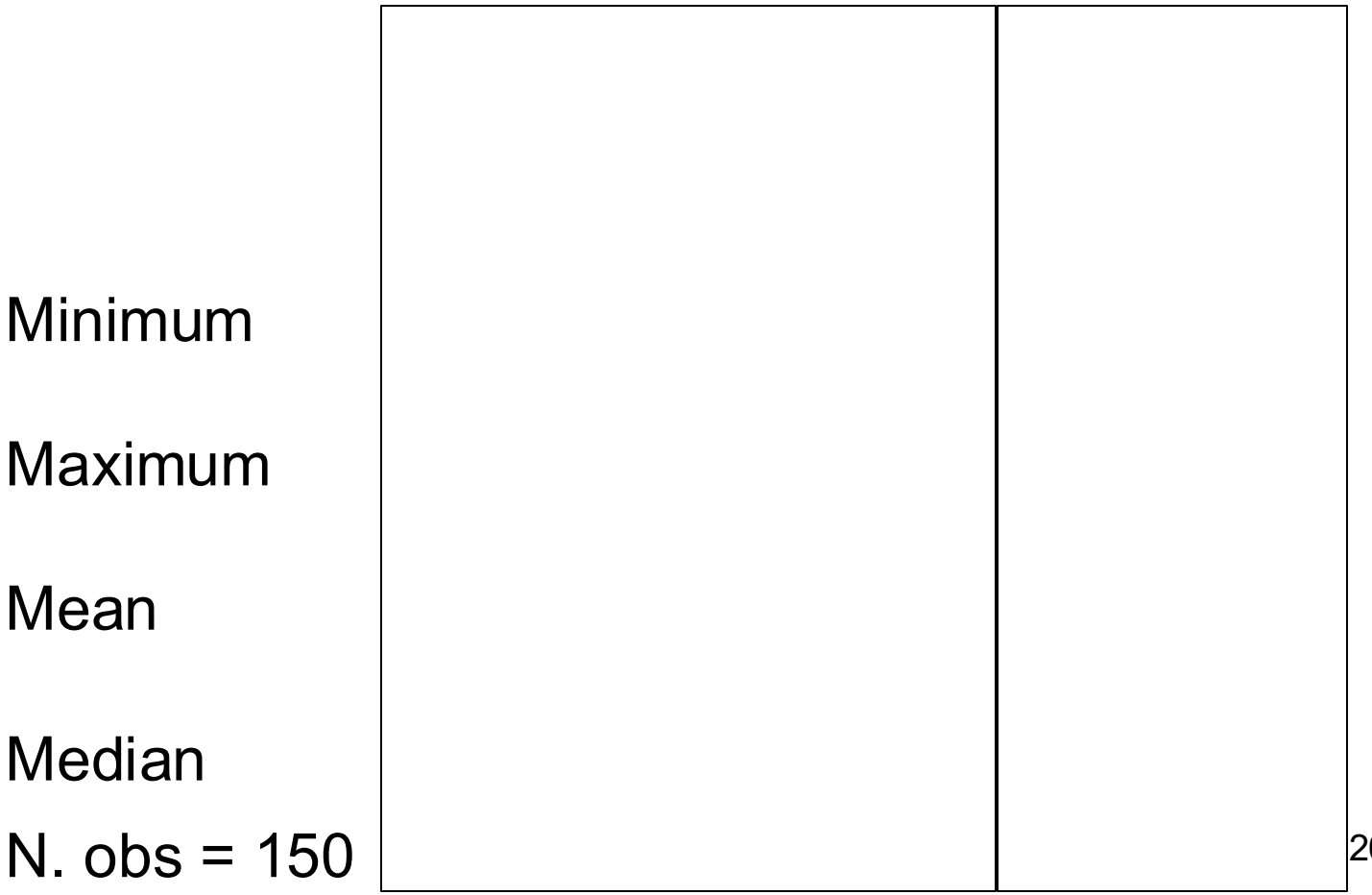


Dealing with dominance

- Can use same approach as frequencies
  redesign, suppress, round etc
- but: best protection is lots of observations
  also deals with the problem of finding it
- consistent with quality again
- dominance is rare: know your data!

SDC example: ranks, maxima, minima

Potential problems with this output?



SDC example: ranks, maxima, minima

- Max and min not always problematic
  assume they are

- Ranks are another form of class disclosure

SDC example: differencing

Potential problems with these tables?

Table 8 Age and socio-economic group

Age bands	socio-econonomic group		
	C2DE	ABC1	Total
50-54	21	11	32
55-59	25	11	36
60-64	28	12	40
65+	31	11	42
	105	45	150

Table 9 Age and socio-economic group

non-diabetics only

Age bands	socio-econonomic group		
	C2DE	ABC1	Total
50-54	17	7	**24**
55-59	19	9	**28**
60-64	23	8	**31**
65+	23	10	**33**
	82	**34**	**116**

- No theoretical solution
- Ad-hoc solution: use higher limits

SDC and statistical quality

- Ideally: no conflict between SDC and research
- Bad for SDC:

  small numbers

  dominant observations/huge outliers

  very skewed distributions

$\Rightarrow$Also to be avoided in analysis

- Be wary of analysis on a single unit

Safe outputs

SDC AND APPLIED RESEARCH

Moving beyond tables…

- Consider
  - linear regression coefficients
  - a scatter plot of regression residuals
  - odds ratio
  - box plots
  - etc…

- do the rules described above apply?
  - do we need to check every statistic?

$\Rightarrow$‘high review’ and ‘low review’ statistics

‘High review’ and ‘low review’ stats

inherently low disclosure risk

↓

‘low review’ statistics

↓

publish

example:
regression coefficients

inherently high disclosure risk

↓

‘high review’ statistics

↓

publish once specific values checked

example:
tables

LRS versus HRS

low review

high review

Classification and clearance

is this output high or low review?

high review

low review

Administrative checks

is this *specific* output okay to release?

yes

no

try again

Send back to researcher to apply SDC

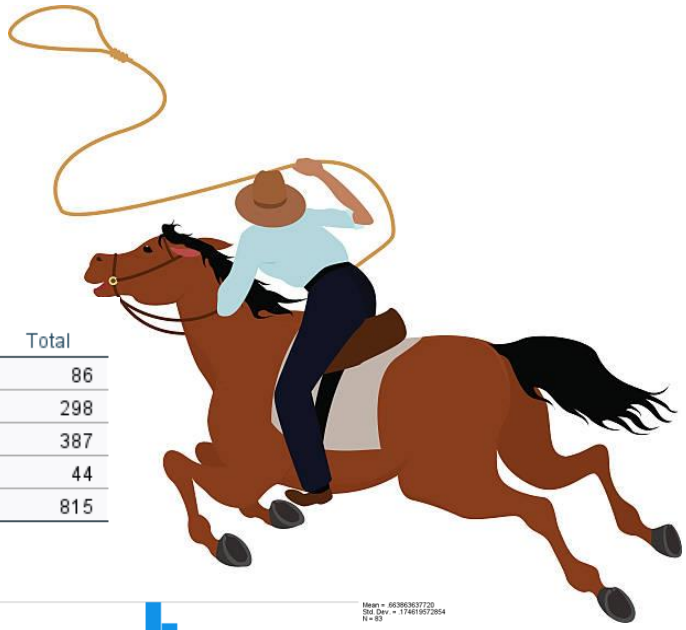
release for publication

Defining LRSs and HRSs

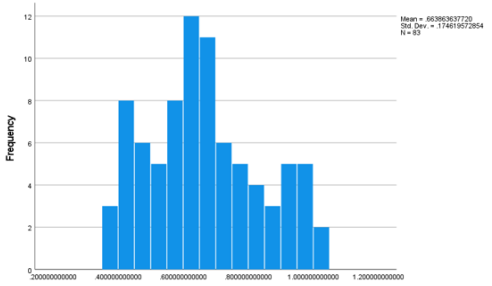
- Are your statistics low or high review?
  - LRS means that you don't need to know about the data
  - some LRSs might have conditions

The statistical barn

- Place homologous statistical analysis into 'statbarns' eg
  - histogram, count table, pie chart
    - ⇒ 'frequencies'



		status		
		dead	successful	Total
grant_type	G	17	69	86
	N	0	298	298
	R	248	139	387
	R/G	0	44	44
Total		265	550	815



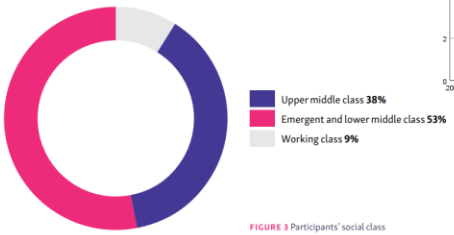


FIGURE 3 Participants' social class

Applying the group rules

- In the Frequencies barn we know all outputs are 

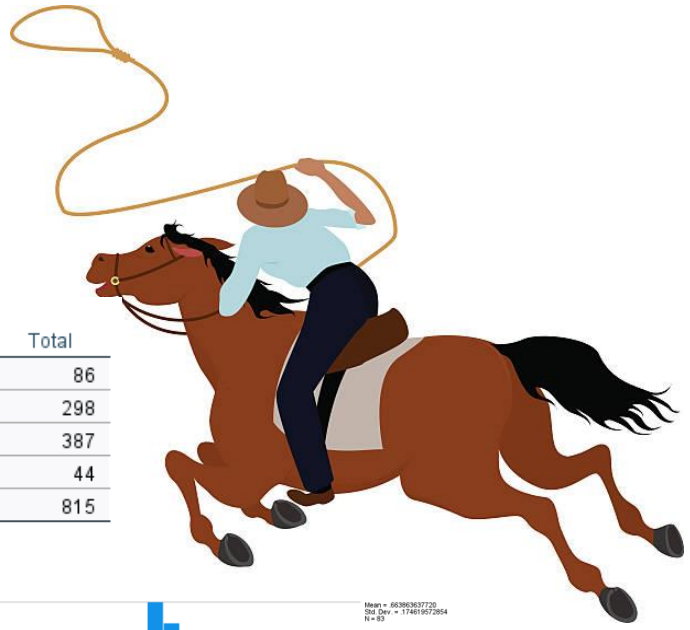
  - With any statistic in the Frequencies barn we need to check:
    - Low counts
    - Differencing
    - Class disclosure

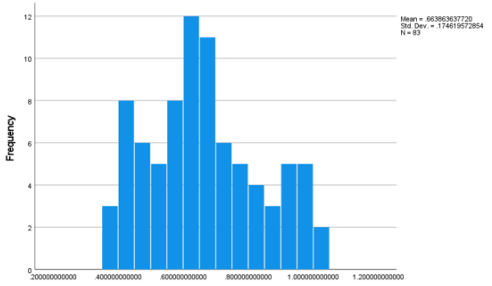
- We would apply the following rules in this barn
  - Minimum count
- Appropriate mitigation techniques for this barn are
  - Cell suppression, noise addition, rounding

The statistical barn

- Place homologous statistical analysis into 'statbarns' eg
  - histogram, count table, pie chart
    - ⇒ 'frequencies'
  - median, interquartile range
    - ⇒ 'position'
  - ANOVA, proportional hazards
    - ⇒ 'correlation'

		status		
		dead	successful	Total
grant_type	G	17	69	86
	N	0	298	298
	R	248	139	387
	R/G	0	44	44
Total		265	550	815





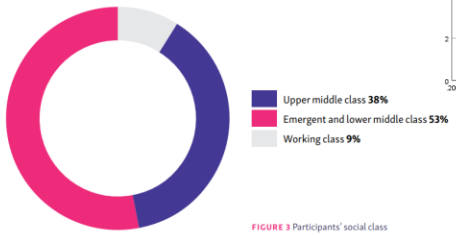


FIGURE 3 Participants' social class

The barns so far

1. Frequencies 

2. ~~Correlation coefficients~~
3. Statistical hypothesis tests
4. Position 

5. End points 

6. Shape
7. Means and total 

8. Mode
9. Non-linear concentration ratios
10. Calculated risk ratios 

11. Hazard/survival tables 

12. Clusters 

13. Linked/multi-level tables 

14. Gini coefficient

- Each barn has its own set of rules for output checking

AI and Machine Learning

- SDC of these types of work is developing area

Work with the support team!

AI and Machine Learning

If developing AI/Machine Learning models to be used outside of the TRE.

- SDC of these types of work is developing area
- Some models cannot be released into the public domain due to privacy risks (e.g. encoding data into the model)

So:

- During application: describe what models and where they're going
- During analysis: consider how you will reduce privacy risk – fortunately also follows good research practice (e.g. restricting hyperparameters to prevent overfitting)
- At output request: provide enough information for an output checker to understand and test the model, and be patient.
- Throughout: Work with the support team

SDC and practical research

- Research outputs have low statistical risk
  - precautionary because we can be
  - apply at the point of release/publication
- SDC aligned with statistical value
  - complex outputs, simple outputs with many obs all low risk
  - be careful output highlighting eg rare events
- be careful with class disclosure