Towards a comprehensive theory and practice of output checking

Ben Derrick

Elizabeth Green

Felix Ritchie

Paul White

A brief history of output SDC

In the beginning

TABLES
late 20th Century

TABLES
everything else
21st Century

SACRO
2023

SACRO Feb-Oct 2023

- Review/revise theory

- Re-integrate theory and practice guidelines

now

- Tools (regular SDC and AI-SDC)

- Community engagement

next

Why a new guideline/manual?

Theory papers

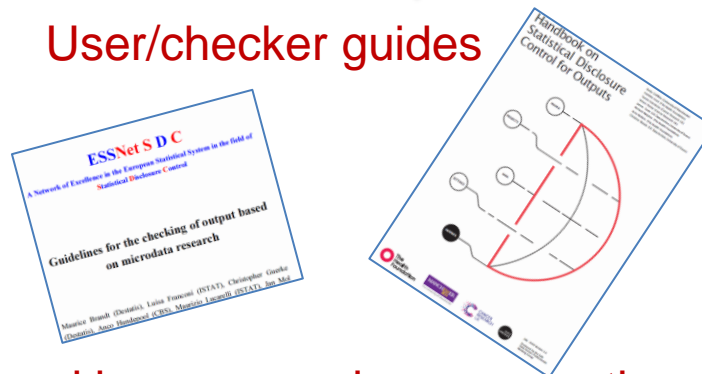Specific, rarely structural

New guide:

Structured

Comprehensive

Definitive

Theoretically sound

Evidenced/sourced

Practice-relevant

but not necessarily
user-friendly

Practice papers

Operational, not integrated

User/checker guides

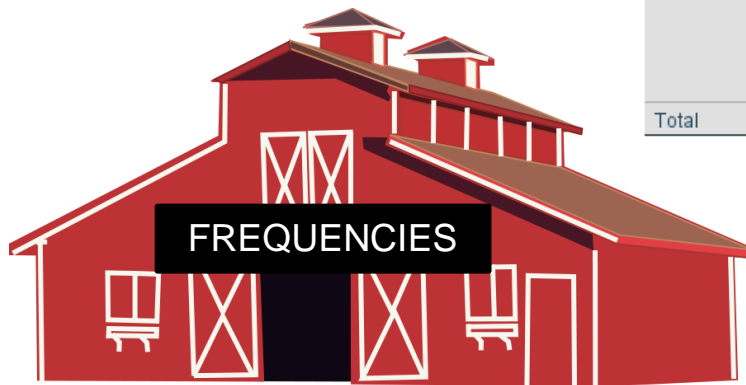General, structured by use, maybe organisation-specific

Structured? okay Comprehensive? No way!

- Building block: 'safe statistic'
  - unsafe: data-dependent; check before release
  - safe: no disclosure risk [minimal check before] release
  - based on mathematical (not statistical) characteristics

- Expand: define all statistics by
  - common disclosure characteristics
  - common mitigation responses

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

FREQUENCIES

Upper middle class **38%**
Emergent and lower middle class **53%**
Working class **9%**

FIGURE 3 Participants' social class

Mean = 663863637720
Std. Dev. = .174619572854
N = 83
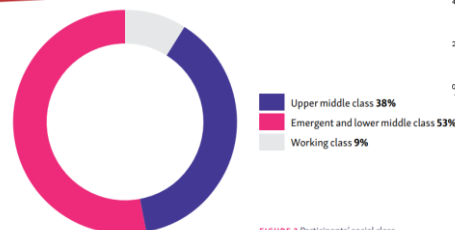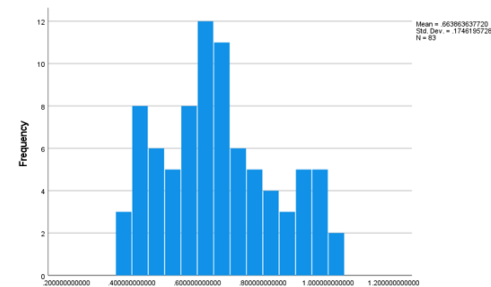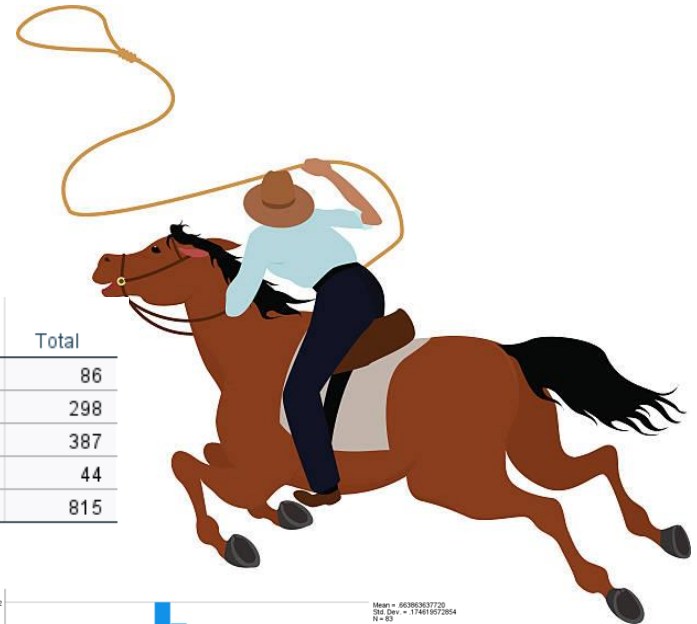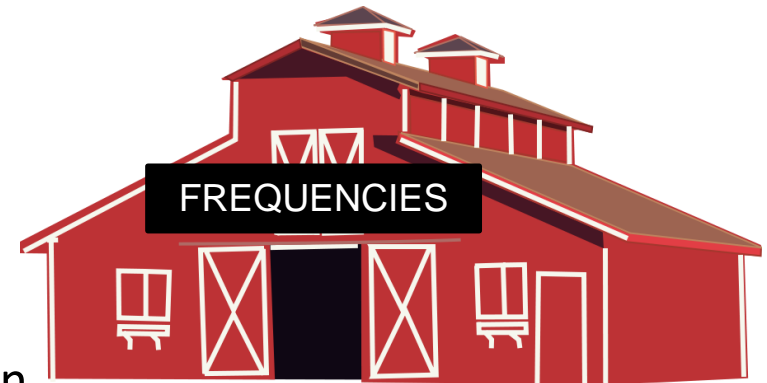
Applying the group rules

- In the Frequencies barn we know all outputs are **UNSAFE**

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

FREQUENCIES

- We would apply the following rules in this barn
    - Minimum count

- Appropriate mitigation techniques for this barn are
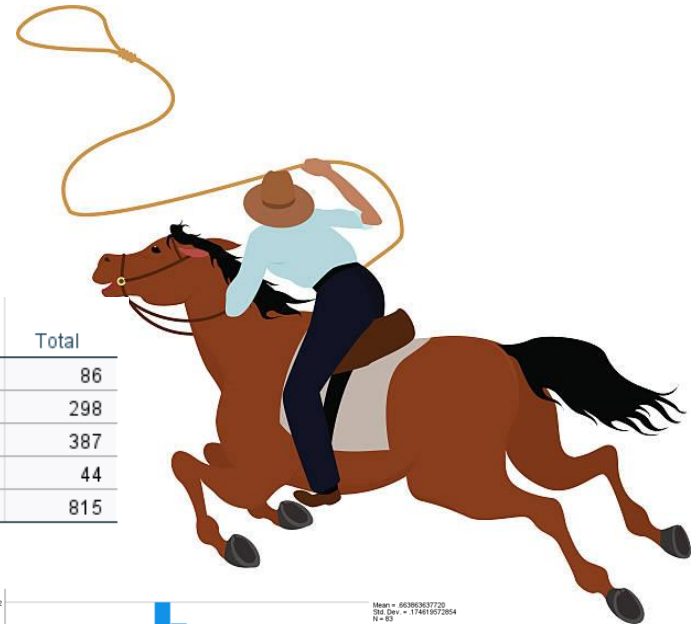   - Cell suppression, noise addition, rounding

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	status 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

FREQUENCIES

Upper middle class **38%**
Emergent and lower middle class **53%**
Working class **9%**

FIGURE 3 Participants' social class

Mean = 663863637720
Std. Dev. = .174619572854
N = 83
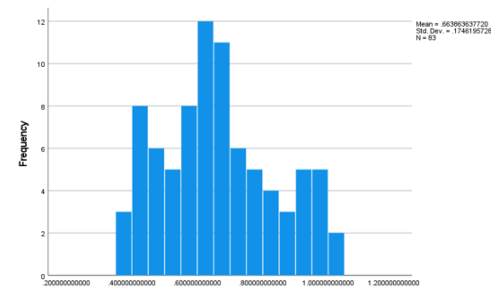
The barns so far

1. Frequencies **UNSAFE**
2. Statistical hypothesis tests ✓
3. Correlation coefficients ✓
4. Position **UNSAFE**
5. End points **UNSAFE**
6. Shape ✓
7. Means and total **UNSAFE**
8. Mode ✓
9. Non-linear concentration ratios ✓
10. Calculated risk ratios **UNSAFE**
11. Hazard/survival tables **UNSAFE**
12. Clusters *TBA*
13. Linked/multi-level tables *TBA*
14. Gini coefficient ✓

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

Progress (as of today)

- Pretty confident on decisions…

- Using in output checker training – positive feedback
- Statbarn model is basis for SACRO (coming up)

- Some rethinking; some revelations

Rethinking example: survival tables

2010 O'Keefe et al JPC

- No detail in tables

- Blur lines in K-M graphs

2019 SDAP manual

- treat as frequency table

- Minimum thresholds
  - tables & graphs

2023 DRAGonTome

- Risk

  1. outliers
  2. absolute dates
  3. differencing via subsets

- Recommendation
  - Approve unless the above
  - Apply threshold to final count

Rethinking example: survival tables

GonTome

2010 O'Ke

- No det
- Blur

...sets

201

...above

...final count

- treat
- Minimum
  - o tables & graphs

mode

regression on single binary variable

linear vs non-linear concentration ratios

Min/max versus medians and percentiles

Revelation example

- What is disclosure?

- Small numbers + finite values = rationale for higher thresholds

- Evidential vs structural zeros

Still to be done

- To be written:
  - Basic concepts
  - Operational issues
  - Directory of other guides
  - FAQs for researchers and output checkers

- Classifications
  - Lookup table – is it comprehensive?
  - Web pages

- Community buy-in
  - **More feedback!**

Thank you; questions?

dragon@uwe.ac.uk
sacro.contact@uwe.ac.uk

Ben.Derrick@uwe.ac.uk
Elizabeth7.Green@uwe.ac.uk
Felix.Ritchie@uwe.ac.uk
Paul.White@uwe.ac.uk

DATA RESEARCH ACCESS & GOVERNANCE NETWORK