**DataLab Safe Researcher training**

Part 3: Safe Outputs and statistical disclosure

control

**Overview**

Part 1 - Working together

Break (10 minutes)

Part 2 - Maintaining data confidentiality

Break (10 minutes)

Part 3 - Statistical disclosure control

• ABS vision

• Shared accountability

• Five Safes Framework

• What is it?

• Why it’s important

• Your role and the ABS’ role

• How might disclosure occur?

• Making outputs safe

• Output Rules

2

**Training outcomes**

**At the end of this training module, you should…**

▪ **Understand** concepts in statistical disclosure

control

▪ **Know** how to prepare safe outputs that are

non-disclosive

**Video: Dr Felix Ritchie**

Associate Professor, University of the West of England

**Why is output checking important?**

• Legal – Only data that is not likely to be identified may

be released

• Departments – Confidence that the data they are

custodian of won’t lead to disclosure

• Providers – People and businesses can have confidence their information is handled appropriately

• Practical – Users don’t need to know everything that

the ABS has released

• But – It’s a shared responsibility

**Collaboration in producing non-disclosive outputs**

• Principles-based output checking

• Follow the DataLab output rules

• Talk with us when you want to break a DataLab

output rule

– Important

– Uncommon

– Non-precedent setting

• Early warning for fast outputs

**Outputs from the DataLab**

▪ Everything is checked by staff in the ABS DataLab

Clearance team before it leaves the DataLab

▪ Safe Outputs

▪ How do we check outputs?

▪ How do you know which outputs are OK?

▪ What do you do if we reject your output?

▪ What statistical disclosure control (SDC) can you do to make your outputs acceptable?

**Statistical disclosure control**

▪ What is SDC?

▪ Checking for disclosure risk in results intended for publication

▪ Mitigating risks by changing the output

▪ Being precautionary

▪ But balancing risk and utility

▪ Consistent with good research

▪ Rule of 10

Each cell should have at least 10 units contributing

Counts of less than 10 should not be able to be worked out

▪ Look at the proposed outputs and we will discuss:

▪ *What if anything in these outputs could be a disclosure risk?*

▪ *What could be done to make it safer (if necessary)?*

▪ *What would you want to tell the output checkers?*

▪ *Any other issues or thoughts?*

**Example 1** – fortnightly income

|  |  |  |
| --- | --- | --- |
|  | Count | % |
| Nil income | 8 | 4.5 |
| $1-$500 | 10 | 5.6 |
| $501-$1000 | 40 | 22.5 |
| $1001-$1500 | 40 | 22.5 |
| $1501-$2000 | 45 | 25.3 |
| $2001-$2500 | 25 | 14.0 |
| $2501 or more | 10 | 5.6 |
| Total | 178 | 100.0 |

**Example 2** – Jobmaker counts over time by age brackets

Age

|  |  |  |  |
| --- | --- | --- | --- |
| Week | 30-35 | 36-40 | Total |
| 1 | 135 | 124 | 259 |
| 2 | 132 | 99 | 231 |
| 3 | 99 | 92 | 191 |
| 4 | 100 | 138 | 237 |
| 5 | 91 | 120 | 211 |
| 6 | 127 | 79 | 206 |

**Example 3** – Count of businesses by rating (score customers give businesses from 1-10)

|  |  |  |  |
| --- | --- | --- | --- |
| Business rating | Count | % | Cumulative % |
| 1 | 12 | 11.65 | 11.65 |
| 2 | 1 | 0.97 | 12.62 |
| 3 | 6 | 5.83 | 18.45 |
| 4 | 14 | 13.59 | 32.04 |
| 5 | 13 | 12.62 | 44.66 |
| 6 | 20 | 19.42 | 64.08 |
| 7 | 1 | 0.97 | 65.05 |
| 8 | 14 | 13.59 | 78.64 |
| 9 | 12 | 11.65 | 90.29 |
| 10 | 10 | 9.71 | 100.00 |

**Example 4** – Jobmaker counts over time by age brackets

Age

|  |  |  |  |
| --- | --- | --- | --- |
| Week | 30-34 | 35-40 | Total |
| 1 | 120 | 139 | 259 |
| 2 | 126 | 105 | 231 |
| 3 | 85 | 106 | 191 |
| 4 | 76 | 162 | 237 |
| 5 | 76 | 135 | 211 |
| 6 | 117 | 89 | 206 |

**Example 2 Example 4**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AgeWeek 30-35 36-40 Total1 135 124 259 |  | AgeWeek 30-34 35-40 Total1 120 139 259 |  |
|  | 2 132 99 231 |  | 2 126 105 231 |  |
|  | 3 99 92 1914 100 138 2375 91 120 2116 127 79 206 |  | 3 85 106 1914 76 162 2375 76 135 2116 117 89 206 |  |

▪ Dominance rules

▪ the largest contributor should not contribute more than

50% to the cell total

▪ the two largest contributors should not contribute more than 67% to the cell total

▪ Look at the proposed outputs and discuss:

▪ *What if anything in these outputs could be a disclosure risk?*

▪ *What could be done to make it safer (if necessary)?*

▪ *What would you want to tell the output checkers?*

▪ *Any other issues or thoughts?*

Example 5 – Profits of pharmacies in local government areas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LGA** | **Profits ($M)** | **Count** | **Top contributor** | **Top two contributors** |
| 1 | 1.65 | 12 | 40% | 76% |
| 2 | 0.94 | 11 | 15% | 29% |
| 3 | 3.22 | 20 | 55% | 65% |
| 4 | 2.1 | 10 | 35% | 57% |
| 5 | 2.05 | 16 | 42% | 56% |
| Total | 9.96 | 69 | 18% | 26% |

Example 7 – Regression coefficients under 4 different models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable**Var1 | **Model 1**1.092\*\*\* (134.76) | **Model 2**0.939\*\*\* (62.36) | **Model 3**0.992\*\*\* (31.53) | **Model 4**0.979\*\*\* (78.23) |
| Var 2 | 0.099\*\*\* (31.07) | 0.147\*\*\* (26.98) | 0.152\*\*\* (10.37) | 0.165\*\*\* (29.79) |
| Var 3 |  | -0.058\*\*\* (-3.79) | 0.060\* (-1.91) | 0.051\*\*\* (3.92) |
| Var 4 |  |  | -0.057 (-1.22) | 0.021\* (1.69) |
| Var 5 |  |  |  | -0.002\*\*\* (-7.27) |
| const | -0.002 (-0.15) | -0.029 (-1.02) | 0.197\*\* (2.03) | 0.102\*\*\* (3.69) |
| N | 74685 | 24432 | 2805 | 23788 |
| r2 | 0.196 | 0.386 | 0.692 | 0.898 |

β-coefficient (t statistics in parenthesis) \* = p<0.1, \*\* = p<0.05, \*\*\* = p<0.01

Example 8 – Characteristics of 2063 households in selected suburb

**Weekly household**

|  |  |  |
| --- | --- | --- |
| **nels Size o**0 | **f household**1 | **income**0 |
| 0 | 1.5 | 1025 |
| 1 | 3 | 2368 |
| 0.7 | 2.9 | 2554 |
| 0.46 | 2.7 | 823 |
| 1 | 5.1 | 3247 |
| 1 | 13 | 4987 |

Min

1st quartile Median Mean

SD

3rd quartile

Max

**Has solar pa**

▪ Combine categories

▪ Round to the nearest 5, 10, 100, 1000, 10000, …

▪ Perturb / add noise

▪ Use words to describe the table

▪ E.g. “*The relative proportions for population X is similar to*

*population Y*”

▪ Suppression (remember secondary)

▪ A note on regressions and modelling

▪ Most of the time regressions are fine. A few things to

be aware of:

▪ Low numbers (ie <10 degrees of freedom)

▪ Removing some records and redoing the regression

▪ High r2 (>0.8 may have disclosure issues)

▪ Categorical variables

▪ Group (class) disclosure

▪ No cells should contain more than 90% of the column or row total

**Award wage**

**Senior**

**Executive**

**teacher**

|  |  |  |
| --- | --- | --- |
| **Age of teacher** | **18-19 year old** | **Junior teacher** |
| 19-20 | 15 | 7 |
| 21-25 | 0 | 34 |
| 26-30 | 0 | 26 |
| 31-35 | 0 | 12 |
| 36-40 | 0 | 0 |

**teacher**

0 0

3 0

11 0

15 8

27 12

**Requesting output clearance**

▪ Help us clear your output quickly by:

➢ Adhering to the output rules

➢ Using your clearance folder

➢ Requesting clearance in a new email chain

➢ Providing detailed descriptions in each field of the output

template

▪ We are human, we make mistakes

▪ Inform us if we have made a mistake in clearing your

output

▪ Don’t use files that have been cleared incorrectly

▪ Delete files and email when requested

▪ Mistakes are investigated for potential breaches and if found to be a breach will be treated accordingly.

**Questions and support**

**DataLab User Guide:**

[https://www.abs.gov.au/statistics/](https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab)

[microdata-tablebuilder/datalab](https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab)

**DataLab contacts:** General enquiries: mailto:data.services@abs.gov.au

**Accessing the ABS DataLab from the User Guide**

