Australian Bureau of Statistics

ABS DATA INTEGRATION & DIGITAL SERVICES

ABS DATA INTEGRATION & DIGITAL SERVICES DataLab

DataLab Safe Researcher Training

Part 3: Safe Outputs and statistical disclosure control

Australian Bureau of Statistics

DataLab

Overview

Part 1 - Working together

(60 minutes)

- ABS vision for the DataLab
- Shared responsibility
- Five Safes Risk Framework

Break (10 minutes)

Part 2 - Maintaining data confidentiality

(40 minutes)

- What does that mean?
- Why is it important?
- Your role and the ABS' role

Break (10 minutes)

Part 3 - Statistical disclosure control

(60 minutes)

- How might disclosure occur?
- Making outputs safe
- Output Rules

Training Outcomes

- Understand concepts in statistical disclosure control

- Know how to prepare safe outputs that are non-disclosive

Statistical Disclosure Control

- **What is SDC?**
  - Checking for disclosure risk in results leaving the 'safe settings'
  - Applying treatments where disclosure risk is too high
- **Principles of SDC**
  - Precautionary
  - Balancing risk and utility
  - Consistent with good research
- **SDC in practice**
  - Output rules – in the User Guide

Why are safe outputs so important

Legal
Only release data that is 'not likely to identify'

Ongoing Data Sharing
Data Custodians have confidence the sharing data won't lead to disclosure

Ongoing Data Collection
People and businesses have confidence their information is handled appropriately

Risk management
Only the results that need to be are removed from the secure environment

Outputs from the DataLab

Everything that leaves the DataLab must first be checked by the ABS DataLab clearance team

Producing safe outputs

- Follow the DataLab output rules
  - Provide evidence
  - Apply treatments

- Principles-based approach to less common analysis

- Requesting exceptions to the standard rules
  - These will be escalated – expect delays
  - You will need to show evidence that it’s important, non-disclosive, and uncommon
  - Any exceptions are non-precedent setting

Main output rules

1. Rule of 10
2. Dominance
3. Model-specific rules

4. Quantiles
5. Group Disclosure
6. Secondary Contributors

Output treatment options

- Treatment should change the output to the point at which is passes the rules
  - Combine categories in tables
  - Round cells to the nearest 5, 10, 100, 1000, 10000, …
  - Perturb/add noise to each cell
  - Use words to describe the output *“The relative proportions for population X is similar to population Y.”*
  - Suppress problematic cells (remember secondary)

Rule of 10

WHY? To prevent the re-identification of units in cells with small counts

WHERE? Rule applies to most outputs (table cells, sums/means, counts used to create charts etc)

Counts of less than 10 should also not be able to be derived from the available data

Each cell should have at least

10 contributing units

Example 1 – Rule of 10

Table: Fortnightly income for persons living on Norfolk Island aged 20-24
Source: Census 2021

	Count	%
Nil income	10	5.6
$1-$500	8	4.5
$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 1 – Rule of 10 TREATED

Table: Fortnightly income for persons living on Norfolk Island aged 20-24
Source: Census 2021

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

	Count	%
Nil income - $500	18	10.1
$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 – Rule of 10

Average Weekly coffees by age group – Persons studying at University

Table 1 – Age groups as per the US Standard

	Age Group		
Coffees per week	<21	21 and over	Total
0	135	124	259
1-2	132	99	231
3-5	99	92	191
6-9	100	138	238
10 or more	91	120	211
Not stated	127	79	206

Table 2 – Age groups as per the Australian Standard

	Age Group		
Coffees per week	<18	18 and over	Total
0	120	139	259
1-2	126	105	231
3-5	85	106	191
6-9	76	162	238
10 or more	76	135	211
Not stated	117	89	206

Example 2 – Rule of 10 TREATED

Average Weekly coffees by age group – Persons studying at University

Table 1 – Age groups as per the US Standard

	Age Group		
Coffees per week	<21	21 and over	Total
0	140	120	260
1-2	130	100	230
3-5	100	90	190
6-9	100	140	240
10 or more	100	120	210
Not stated	130	80	210

Table 2 – Age groups as per the Australian Standard

	Age Group		
Coffees per week	<18	18 and over	Total
0	120	140	260
1-2	130	110	230
3-5	90	110	190
6-9	90	160	240
10 or more	80	140	210
Not stated	120	90	210

Dominance

WHY? To prevent the re-identification of units that contribute a large percentage of a cell's total value

WHERE? Applies mainly to sums/totals and means

The **largest** contributor must contribute less than 50%

The **two largest** contributors must contribute less than 67%

Example 3 - Dominance

Total turnover ($M) of all pharmacies by Local Government Area

LGA Code	Total Turnover	No. of Businesses
1	1.65	12
2	0.94	11
3	3.22	20
4	2.10	10
5	2.05	16
Total	9.96	69

Example 3 - Dominance

Total turnover ($M) of all pharmacies by Local Government Area

LGA Code	Total Turnover	No. of Businesses	Turnover of largest business	Turnover of 2nd largest business	Proportion from largest business to total	Proportion from largest two businesses to total
1	1.65	12	0.66	0.59	40%	76%
2	0.94	11	0.14	0.13	15%	29%
3	3.22	20	1.77	0.32	55%	65%
4	2.10	10	0.74	0.46	35%	57%
5	2.05	16	0.86	0.29	42%	56%
Total	9.96	69	1.77	0.86	18%	26%

Example 3 – Dominance - TREATED

Total turnover ($M) of all pharmacies by Local Government Area

LGA Code	Total Turnover	No. of Businesses	Turnover of largest business	Turnover of 2nd largest business	Proportion from largest business	Proportion from largest two businesses
1&3	4.87	32	1.77	0.66	36%	50%
2	0.94	11	0.14	0.13	15%	29%
4	2.1	10	0.74	0.46	35%	57%
5	2.05	16	0.86	0.29	42%	56%
Total	9.96	69	1.79	0.8	18%	26%

OR

"Total turnover ranking for the five LGAs of interest were (from largest to smallest): LGA 3, 4, 5, 1 and then 2."

Model-specific rules

WHY? Designed to prevent the re-identification of units using overfitted models and/or residuals

WHERE? All modelling outputs

The model should have at least **10 degrees of freedom**

The **R-squared** for least squares regression should be <= 0.9

Individual residuals cannot leave the DataLab

Extra rules when the independent variables are all categorical (contact the ABS)

Example 4 – Model-specific rules

Linear regression that looks at personal income as a function of a range of variables.

Variable	Model 1	Model 2	Model 3	Model 4
Sex	11.34	8.35	8.12	8.33
Age	1.61	1.56	1.55	1.55
SEIFA (index value)		17.28	17.33	17.33
Completed Yr 12			-6.76	-7.93
Has Bachelor Degree				2.36
Constant	36.85	-9.88	-5.23	-5.27
N	371	371	371	371
r^2	0.23	0.78	0.79	0.97

Minimum contributors for quantiles

WHY? To prevent the re-identification of units in from a group with small counts

WHERE? Any quantiles, maximum, minimum, range

Each "bin" must have **at least 5 contributors**

No **minimums** or **maximums** out of DataLab

	Minimum contributors
Percentiles	500
Deciles	50
Quartiles	20
Median	10

Example 5 - Quantiles

Age	Count
0	11
1	0
2	4
3	6
4	14
5	17
6	11
7	17
8	9
9	6
10	2
11	1
12	0
13	0
14	2
Total	**100**

	Original	Requirement	Treated
Minimum	0	Min 10 in cell	OK - 0
5th percentile	0	100 total contributors	OK - 0
Median	5	10 total contributors	OK - 5
95th percentile	9.5	100 total contributors	OK – 9.5
99th percentile	14	500 total contributors	Cannot clear
Maximum	14	Min 10 in cell	Cannot clear

Group Disclosure Rule

WHY? To protect the disclosure of a previously unknown attribute of an individual or business from a given group, where that group has a common feature

WHERE? Totals, means, proportions, counts

Particularly important where there is a risk of adverse consequences to the group

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

Example 6 – Group disclosure

Whether ever incarcerated, by selected occupations

	Ever incarcerated (No.)		Ever incarcerated (Row %)	
Occupation Code	Yes	No	Yes	No
Plumber	12	200	6%	94%
Sales Assistant	110	102	52%	48%
Police officer	0	36	0%	100%
Librarian	140	11	93%	7%

Secondary contributor rules

WHY? Designed to protect the confidentiality where data has been collected and output about one unit (primary contributor) but could disclose information about a higher-level unit (secondary contributor)

WHERE? Output from multi-level datasets

At least 5 businesses or 10 households

In addition to the Rule of 10 for the primary contributor

Example 7 – Secondary contributors

- Number of persons per SA3 working full time in the mining industry
- Source: Employee, Earnings and Hours Survey

Area	Total Employees (weighted)		
North	10,345		
South	5,023		
East	44,553		
West	24,344		
Mid	701		

Example 7 – Secondary contributors

- Number of persons per SA3 working full time in the mining industry
- Source: Employee, Earnings and Hours Survey

Area	Total Employees (weighted)	Total Persons (unweighted)	Number of unique Businesses
North	10,345	1057	7
South	5,023	543	2
East	44,553	4754	13
West	24,344	2489	12
Mid	701	65	1

Other outputs

- Charts/graphs – supply underlying counts
- Indexes – Explain index construction
- Code – remove counts and other data

Help us to clear to your outputs quickly

- **Checking** your output meets the rules and applying treatments
- Clearly **labelling** and **formatting** your output
- Providing the required **supporting data**
- Copying both outputs and evidence to your **O:/Output drive**

Help us to clear to your outputs quickly

- To request clearances, use the clearance request tile in the myDATA portal
- Providing detailed **descriptions** in each field

Do not put counts or other data into emails

Outputs from the DataLab

- 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 emails when requested
- Mistakes are investigated for potential breaches and if found to be a breach will be treated accordingly.

Questions and support

Use information on the ABS website

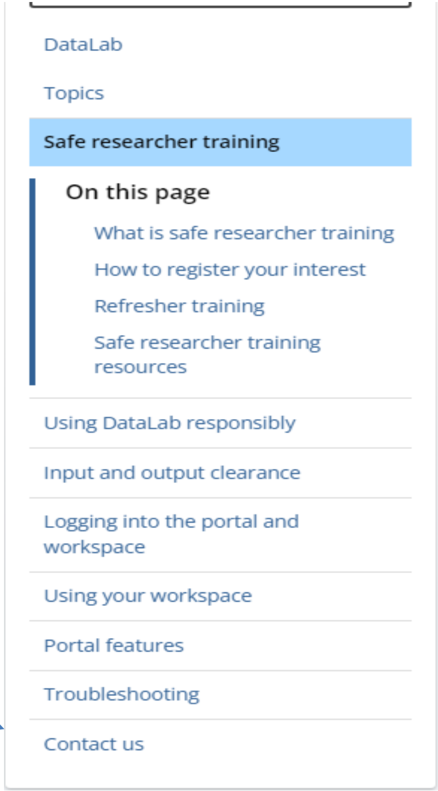
There are rules, and examples plus this learning material.

DataLab User Guide

https://www.abs.gov.au/statistics/microdata-tablebuilder/datalab

DataLab enquiries

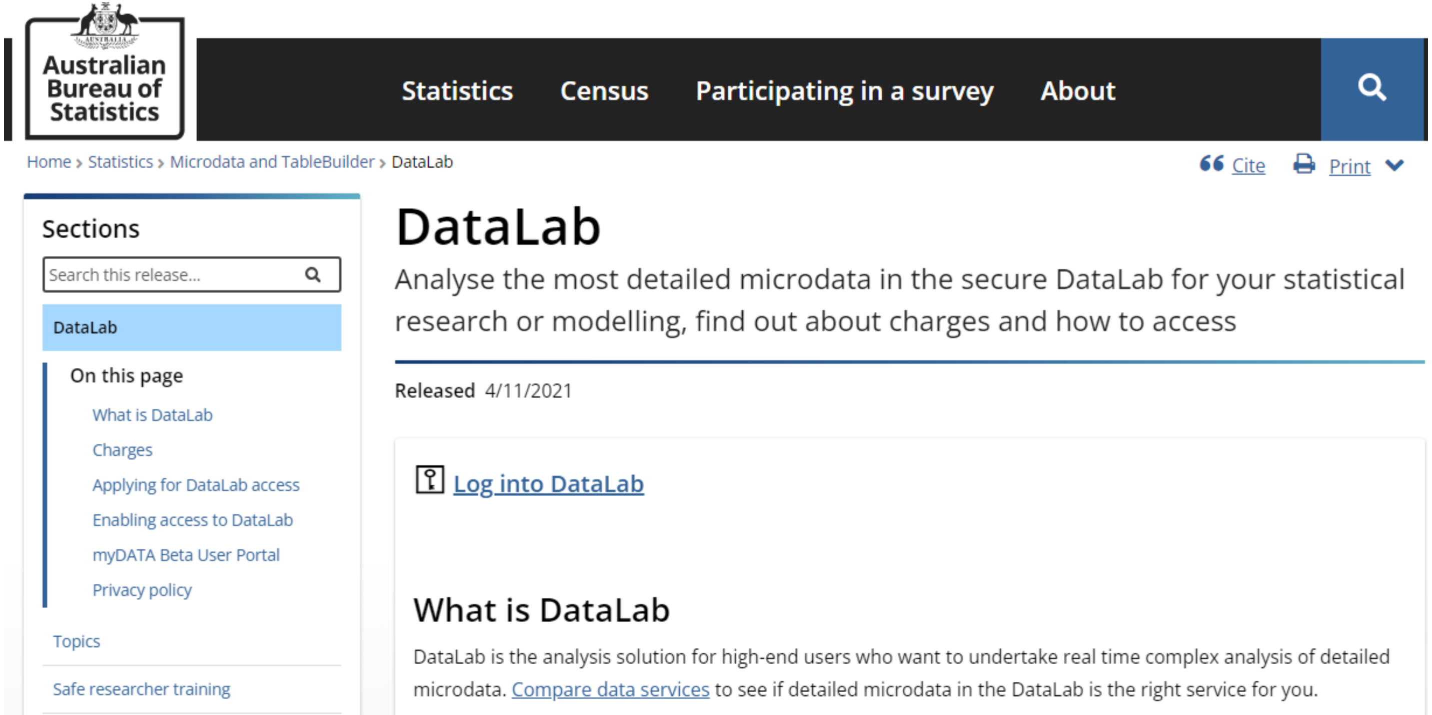
Go to “Contact us” in the user guide and choose the template that matches your query


DataLab
Topics
Safe researcher training
On this page
What is safe researcher training
How to register your interest
Refresher training
Safe researcher training resources
Using DataLab responsibly
Input and output clearance
Logging into the portal and workspace
Using your workspace
Portal features
Troubleshooting
Contact us

What’s next …… myDATA portal

- Login to the myDATA portal and download the quiz and all the forms
  - Complete the quiz (within 3 months from the training date), and read, sign and submit all the forms via email:
    - to: info@mydata.abs.gov.au
    - subject line: DataLab training quiz and forms

Accessing the DataLab from the User Guide



DataLab

Analyse the most detailed microdata in the secure DataLab for your statistical research or modelling, find out about charges and how to access

Released 4/11/2021

Log into DataLab

What is DataLab

DataLab is the analysis solution for high-end users who want to undertake real time complex analysis of detailed microdata. Compare data services to see if detailed microdata in the DataLab is the right service for you.

Australian Bureau of Statistics

ABS DATA INTEGRATION & DIGITAL SERVICES

ABS DATA INTEGRATION & DIGITAL SERVICES

DataLab

DataLab Safe Researcher Training

Thank you for attending today’s training