Australian Bureau of Statistics

ABS DATA INTEGRATION & DIGITAL SERVICES

ABS DATA INTEGRATION & DIGITAL SERVICES

DataLab

DataLab Safe Researcher Training

Module 3: DataLab Output Clearance Rules and Safe Outputs

Training Outcomes

Understand the output clearance rules

Know how to apply these rules to produce safe outputs

Australian Bureau of Statistics | DataLab

Any information from the DataLab must be submitted for clearance before use outside the DataLab

How to produce safe outputs

Follow the DataLab output rules

Apply treatments if the rules fail

Provide supporting evidence

Standard output rules

- Rule of 10
- Differencing
- Dominance
- Quantiles
- Regressions (OLS)
- Group disclosure
- Secondary contributors

Output treatment options

Treatment should change the output to the point at which it passes the rules.

- Combine categories in tables
- Suppress problematic cells
  - Apply secondary suppression, or
  - Apply rounding: after suppression of counts less than 10, round ALL remaining counts to the nearest 10 or randomly up and down to base 5.
- Use words to describe the output “The relative proportions for population X are similar to population Y.”

More details and examples of using these are available on the DataLab Clearance webpage.

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.)

Each cell should have at least **10 contributing units**

Example 1 – Rule of 10

Table: Fortnightly income for persons living in an SA2 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 in an SA2 aged 20-24
Source: Census 2021

	Count	%
Nil income	10	5.6
$1-$500	SUPPRESSED	SUPPRESSED
$501-$1000	40	22.5
$1001-$1500	40	22.5
$1501-$2000	45	25.3
$2001-$2500	25	14.0
$2501 or more	SUPPRESSED	SUPPRESSED
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

Differencing

When the difference between reported counts is also a count of people or businesses, the rule of 10 applies.

Check differencing for subset relationships, e.g. in:

- Overlapping variables or groups
- Varying levels of aggregation

Counts of less than **10** cannot be **derived** from the available data

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	90	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	80	160	240
10 or more	80	140	210
Not stated	120	90	210

Dominance

WHY? To prevent attribute disclosure 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 **50% or less**

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

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."

Regressions (Ordinary Least Squares)

WHY? Designed to prevent attribute disclosure of units using overfitted models and/or residuals

WHERE? All OLS outputs

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

The **R-squared** should be ≤ 0.9

Individual residuals cannot leave the DataLab

When **all** independent variables are **categorical**, provide evidence that category means meet the rule of 10 and dominance (or suppress intercept)

Example 4 – OLS regression

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

Quantile rules

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

WHERE? Any quantiles, maximum, minimum, range

Each “bin” must have **at least 5** contributors

Minimums and **maximums** are generally not safe

	Minimum contributors
Percentiles	500
Vigintiles	100
Deciles	50
Quartiles	20
Median	10

Example 5 - Quantiles

Distribution of ages of children with a disability in an SA2

Statistic	Value	Evidence	Rule	Outcome
Minimum	0	11	Rule of 10	Passed
Maximum	14	2	Rule of 10	Failed – suppress
Median	5	100	10 total contributors	Passed
5th percentile	0	100	100 total contributors	Passed
95th percentile	9.5	100	100 total contributors	Passed
99th percentile	14	100	500 total contributors	Failed – suppress

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? Counts, proportions, percentages, totals

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 type of 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

- Other models
- Charts/graphs – supply underlying data
- Code – remove counts and other data

How to prepare outputs for efficient clearance

- **Check** your output meets the rules and apply treatments
- Clearly **label** and **format** your output
- Provide the required **supporting evidence**
- Copy both outputs and evidence to your **O:/Output drive**

How to request output clearance

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

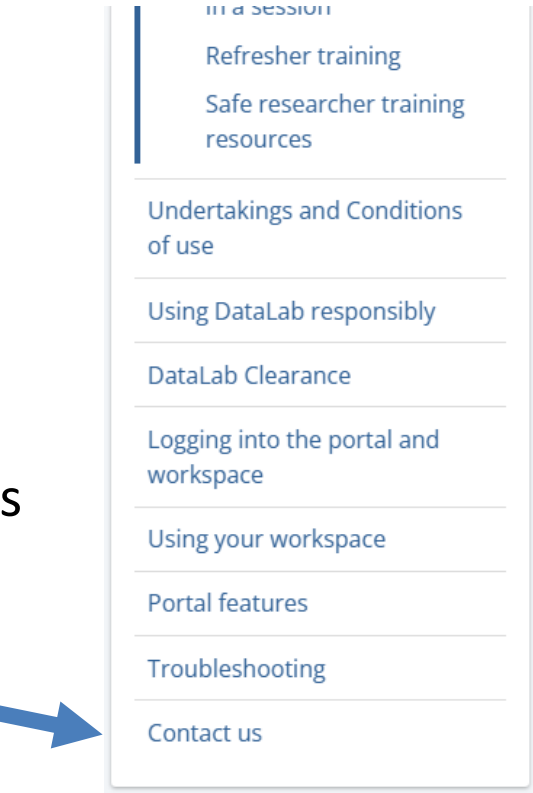
Do not put counts or other data into myDATA or emails

Questions and support

Use information on the ABS website:

- Rules and examples: **DataLab Clearance**
- **DataLab User Guide**

DataLab enquiries: choose the template that matches your query from the **Contact us** web page

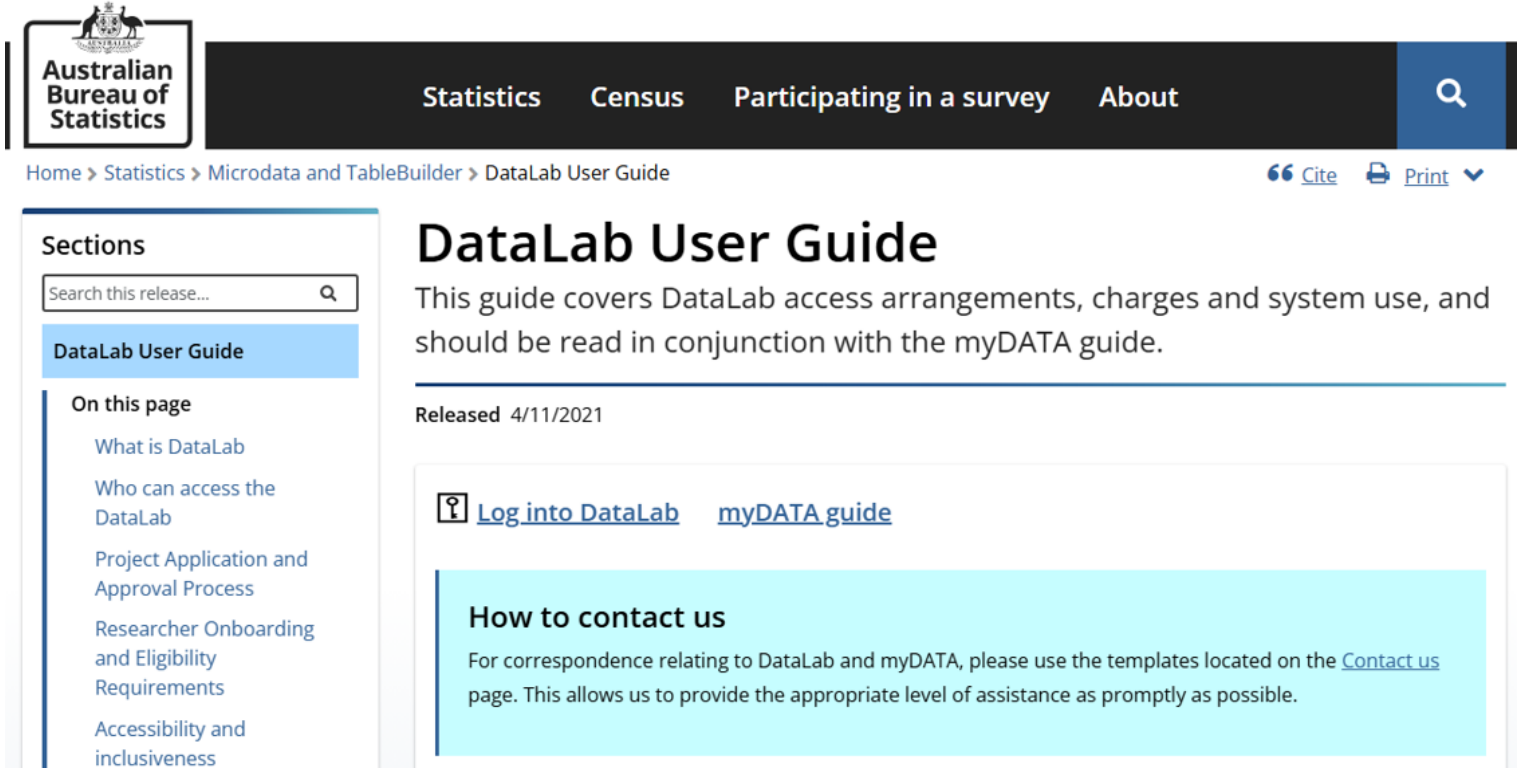


What's next … myDATA portal

Login to the myDATA portal and download the quiz and all the undertakings and conditions of use forms

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

Accessing the DataLab from the User Guide



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

For all queries, please use the correct template from our Contact Us page