

ACRO Guide for researchers

Contents

1. Overview	2
2. Statistical disclosure rules used	2
3. Output generated	3
4. Using ACRO – steps for the researcher to take.....	10
4.1 Findings files and checking the SDC rules	11
4.2 Set-up.....	11
4.3 Sending outputs for checking	12
4.4 Finalising	13
5. Hints	13

March 2021

Version 01.j (pilot)

Authors

Felix Ritchie, Elizabeth Green and Jim Smith, University of the West of England, Bristol.

For queries, please contact felix.ritchie@uwe.ac.uk.

1. Overview

ACRO (Automated disclosure Control of Research Outputs) is designed to reduce the burden of checking for confidentiality risks in the outputs of researchers and analysts in a secure environment.

Researchers work as usual, but use the word 'safe' to prefix results they want to have exported from the research environment. This creates a copy of the results in an Excel workbook, automatically carries out disclosure checks, and sends a message to the researcher about the result of the checks.

If the output passes the disclosure control tests, or an exception is requested (see below) then the full requested output is included in the spreadsheet. Output which is rejected (either automatically or after manual review) is not included on the spreadsheet. The researcher sees immediately which output is automatically refused and can update the results as necessary.

Researchers can request for unacceptable values to be automatically suppressed. They can also request that an 'exception' be granted; that is, a request would normally fail the checks but there might be reasons why the output can be released, such as the transformation of the data (or publication of weighted results only with high weights).

When the researcher has completed her work, then the workbook can be sent to the output checking team, who will review the exception requests and release (or not) the output to the researcher.

The disclosure checking process operates in the background and does not affect the normal display of results. However, only results saved to the workbook by ACRO are part of the review process. Results which cannot be managed through this process need to be submitted for review through the normal channels.

The pilot version is currently coded for Stata only. ACRO currently is coded for:

- Tabular outputs: table, tab or tabulate
- Estimation: regress, xtreg, probit, logit, and all the associated statistics such R^2 , t-stats
- Graphs: all graph commands, exported as images (not data)

2. Statistical disclosure rules used

The following statistical disclosure control (SDC) rules are applied. Not all rules apply in all cases, or to all datasets. Your system manager will be able to tell you which. This is the default set.

Rule	Description	Applies to	Relevant commands (as currently implemented)
Threshold	Minimum number of observations underlying a statistic	All linear statistics: frequencies, mean, median, sums etc	table tabulate
N-K	The N largest observations should not count for more than K% of the total	As for threshold, but doesn't apply to frequencies	table tabulate
P-ratio		As for N-K rule	table tabulate
Table rule	SDC rules are applied to each table cell independently; any cell can pass or fail	Applies to all tables	table tabulate

Maximum & minimum	Not allowed	Any magnitude	table tabulate
Degrees of freedom	Analytical outputs must have at least K degrees of freedom	Analytical results, including estimation and testing	regress xtreg logit probit test ttest

These rules are of necessity crude, but in general these should cause few problems for research outputs. SDC is generally consistent with good research: large numbers of observations, no extreme outliers, well-behaved distributions and so on, all lead to both quality analysis and low disclosure risk. However, because ACRO applies these rules irrespective of the sensitivity of the data, it can over-protect the data; for example if your dataset covers the working population, there is no disclosure risk in noting that the minimum and maximum ages are 16 and 65. For this reason, ACRO has a built in mechanism to allow for an *exception* to the rules to be requested where

- there is no disclosure risk, and
- it is important, and
- it is genuinely an exception.

3. Output generated

Results are sent to an Excel macro-enabled workbook (the macro allows the output-checker to automatically clean and format the file).

The first page of the workbook lists the outputs generated, the results of the assessment, and additional information such as the type of output and whether any exception was requested. For example, this is the output from running **test_file_small.do**:

Sheet	Automatic check	Final decision	Description safe/unsafe	Reason for automatic decision	Exception request	Additional notes
activity	ok	ok	unsafe statistic: table	pass	n/a	
graph_test	review		graph: twoway	review required		
max_act	ok	ok	unsafe statistic: table	fail; suppression app	n/a	
output_1	ok	ok	unsafe statistic: tabulate	pass	n/a	
output_2 A	fail	fail	unsafe statistic: tabulate	fail	n/a	
output_2 B	ok	ok	unsafe statistic: tabulate	pass	n/a	
output_2 C	ok	ok	unsafe statistic: tabulate	pass	n/a	
output_2 D	fail	fail	unsafe statistic: tabulate	fail	n/a	
output_3	fail	fail	unsafe statistic: table	fail	n/a	
output_4	review		unsafe statistic: table	fail; exception request	trust me, I'm a professor	
output_5	ok	ok	safe statistic: regress	pass	n/a	
small_act A	review		unsafe statistic: table	fail; exception request	It's not feasible to identify the charities from this information	
small_act B	review		unsafe statistic: table	fail; exception request	It's not feasible to identify the charities from this information	
small_act C	review		unsafe statistic: table	fail; exception request	It's not feasible to identify the charities from this information	

Figure 1 Description page of the workbook

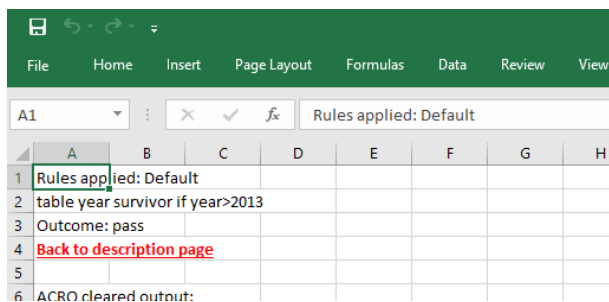
The first column contains a hyperlink to simplify navigation. It can be seen here that:

- the sheets “activity” and ‘output_1’ passed the automatic check; although ‘unsafe’ statistics, they have passed threshold and dominance checks appropriate for the ‘table’ or ‘tabulate’ commands
- ‘graph_test’ has been set for review as this cannot be done automatically at present

- the sheet “max_act” failed, but suppression was applied, so it is ok to release
- the sheet “output_3” failed the SDC checks
- sheets “output_4” and “small_act” failed, but a reason has been given as to why each is an exception; these will be manually reviewed by the output checker

On the other spreadsheets in the workbook, the detailed results are stored. The first four lines show

- The SDC rules being applied
- The command that was run
- The result of ACRO’s checks
- A link back to the front page



	A	B	C	D	E	F	G	H
1	Rules applied: Default							
2	table year survivor if year>2013							
3	Outcome: pass							
4	Back to description page							
5								
6	ACRO cleared output:							

Figure 2 Output sheet contents list

The contents of the rest of the page depend on the check result. In running its checks, ACRO runs its own version of the command (for example, for tabulations it removes weights, ignores irrelevant summary statistics, and forces frequencies to be calculated). We refer to this as the SDC-summary.

If the check was **pass**, **fail-but-exception-requested** or **review_required** (the last is always the outcome for graphs in this version) then the sheet shows

- The SDC-summary in full detail, and
- The output that would have been produced by Stata

If the check was **fail-but-apply-suppression** then the sheet shows

- The SDC-summary with appropriate values suppressed

The original Stata output is not presented, as it will have suppressed values in it.

If the check was **fail** with no suppression or request for an exception then no output is presented. The sheet just shows

- The set of SDC rules being applied
- The command that was run
- A statement that it failed the SDC tests

If the test files are run, the “activity” spreadsheet shows:

1	Rules applied: Default							
2	table year survivor if year>2013 , contents(freq mean inc_activity sd inc_activity)							
3	Outcome: pass							
4	Back to description page							
5								
6	ACRO cleared output:							
7	year	survivor	frequency	freq_3_in	sd_inc_act	freq_2_in	mean_inc	problems
8	2014	Dead in 20	50	46	1701519	46	672290.8	ok
9	2014	Alive in 20	103	103	51898800	103	16654627	ok
10	2015	Dead in 20	50	31	1515565	31	485240.2	ok
11	2015	Alive in 20	103	98	55123052	98	16599920	ok
12	ACRO validated output							
13								
14	survivor							
15	year		Dead in 2015	Alive in 2015				
16								
17	2014		50	103				
18			672290.8	1.67e+07				
19			1701519	5.19e+07				
20								
21	2015		50	103				
22			485240.2	1.66e+07				
23								

Figure 3 A successful tabulation request

Figure 3 shows an extract from the sheet “activity”, the result of running the command to show the mean and standard deviation of the variable `inc_activity` for combinations of the year/survivor categories. The command is

```
table year survivor if year>2013 , contents(freq mean  
inc_activity sd inc_activity)
```

The first row shows that the default (ie not dataset-specific) rules are being applied. The command is listed in the second row of the sheet. The third row shows that there are no SDC issues with this output.

Lines 6-11 show the SDC-summary. The categories of the tabulation (in this case, ‘year’ and ‘survivor’ are in the first columns. The first column after the categories is the frequency of non-missing responses for those category values. This must exceed the relevant threshold to pass the SDC check. After the frequency, any other statistic requested are presented, preceded by their respective frequencies. This second frequency is not necessarily the same as the cell frequency as it depends upon non-missing values. For example, cell C8 shows that there are 50 observations in 2014 classed as ‘dead’. However, cells D8 and F8 show that there are only 46 non-missing values of `inc_activity` used to calculate the standard deviation or mean of `inc_activity`.

Lines 12-23 show the output as Stata would have produced it. This can be reformatted by the researchers using Excel’s Text-to-Columns command.

	A	B	C	D	E	F	G	H	I	J
1	Rules applied: Default									
2	table year survivor if year>2013 , contents(freq mean inc_activity max inc_activity)									
3	Outcome: fail; suppression applied									
4	Back to description page									
5										
6	ACRO cleared output:									
7	year	survivor	frequency	freq_3_in	max_inc	freq_2_in	mean_inc	problems		
8	2014	Dead in 20	50	46	n/a	46	672290.8	max/min not allowed;		
9	2014	Alive in 20	103	103	n/a	103	16654627	max/min not allowed;		
10	2015	Dead in 20	50	31	n/a	31	485240.2	max/min not allowed;		
11	2015	Alive in 20	103	98	n/a	98	16599920	max/min not allowed;		
12										
13										

Figure 4 Output with suppression applied

Figure 4 shows output which failed the tests, but where suppression has been applied (taken from the spreadsheet “max_act”). In contrast to the first example, here the maximum value of inc_activity and the mean has been requested. As the maximum is not allowed, this would normally fail the check. However, the researcher selected the ‘suppress’ option. We can therefore produce a ‘safe’ output with problematic values replaced with “n/a”

Obviously, in this case the unadjusted Stata output is not presented, as it would breach the rules.

	A	B	C	D	E	F	G
1	Rules applied: Default						
2	table year survivor , by(grant_type)						
3	Outcome: fail						
4	Back to description page						
5							
6	*** no output cleared ***						
7							
8							
9							

Figure 5 Failed output

Figure 5 shows an output which has failed to pass the SDC tests, and no exception or suppression was requested (spreadsheet “output_3”. The sheet merely recognises that this command was run.

	A	B	C	D	E	F	G	H	I
1	Rules applied: Default								
2	table year survivor , by(grant_type)								
3	fail; exception requested Justification: trust me, I'm a professor								
4	Back to description page								
5									
6	ACRO cleared output:								
7	year	survivor	grant_type	frequency	problems				
8	2010	Dead in 20	G	3	below threshold;				
9	2010	Dead in 20	R	47	ok				
10	2010	Alive in 20	G	12	ok				
11	2010	Alive in 20	N	59	ok				
12	2010	Alive in 20	R	24	ok				
13	2010	Alive in 20	R/G	8	below threshold;				
14	2011	Dead in 20	G	3	below threshold;				

Figure 6 Output requesting an exception

In the case shown in figure 6 the output has failed because of frequencies below the threshold. However, as the researcher has asked for an exception, given in cell A3 (“Trust me, I’m a professor”). The full data is therefore made available, including the original Stata output from row 44 downwards (not shown), so that the output checker can make a judgement.

If the output checker rejects the researcher’s case for an exception, which in this case she would, then before releasing the spreadsheet, she runs a macro which blanks out everything apart from the first two lines and a note that the request was denied. If however the output checker agrees with the exception, then the full output is released.

If the user runs the commands on subsets of the data by using “**by** ... : ” at the start of the command, the results are shown on different sheets with _A, _B, _C etc added to the sheet name, as in the case of the output ‘small_act’ below:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Rules applied: Default											
2	table year survivor if year>2013 , contents(freq mean inc_activity max inc_activity) By grant_type: [grant_type=G]											
3	fail; exception requested Justification: It's not feasible to identify the charities from this information											
4	Back to description page											
5												
6	ACRO cleared output:											
7	year	survivor	frequency	freq_3_in	max_inc	freq_2_in	mean_inc	problems				
8	2014	Dead in 20	3	3	137000	3	60063.33	below threshold; max/min not allowed;				
9	2014	Alive in 20	12	12	1.69E+08	12	41632016	max/min not allowed;				
10	2015	Dead in 20	3	3	115000	3	51454	below threshold; max/min not allowed;				
11	2015	Alive in 20	12	9	1.7E+08	9	52462600	below threshold; max/min not allowed;				
12	ACRO validated output											
13												
14		survivor										
15	year		Dead in 2015		Alive in 2015							

Figure 7 Labelling of sheets for subsets of the data

The description of the top of the page shows the particular value being used at each point. For example

	A	B	C	D	E	F	G	H	I	J	K	L
1	Rules applied: Default											
2	table year survivor if year>2013 , contents(freq mean inc_activity max inc_activity) By grant_type: [grant_type=G]											
3	fail; exception requested Justification: It's not feasible to identify the charities from this information											
4	Back to description page											
5												
6	ACRO cleared output:											
7	year	survivor	frequency	freq_3_in	max_inc	freq_2_in	mean_inc	problems				
8	2014	Dead in 20	3	3	137000	3	60063.33	below threshold; max/min not allowed;				
9	2014	Alive in 20	12	12	1.69E+08	12	41632016	max/min not allowed;				
10	2015	Dead in 20	3	3	115000	3	51454	below threshold; max/min not allowed;				
11	2015	Alive in 20	12	9	1.7E+08	9	52462600	below threshold; max/min not allowed;				
12	ACRO validated output											
13												
14		survivor										
15	year		Dead in 2015		Alive in 2015							

Figure 8 identifying the by-values

It can be seen (line 2) that this output is produced by the subgroup where 'grant_type' is equal to "G".

Finally, it should be noted that SDC checks take place on unweighted data; that is the frequency count is checked before weights are applied. The SDC summary at the top of the sheet lists unweighted frequencies; the full Stata output below has the weighted values.

1	Rules applied: Default				
2	table year survivor [pweight=wgt] ,				
3	Outcome: pass				
4	Back to description page				
5					
6	ACRO cleared output:				
7	year survivor frequency problems				
8	2010 Dead in 2015 50 ok				
9	2010 Alive in 2015 103 ok				
10	2011 Dead in 2015 50 ok				
11	2011 Alive in 2015 103 ok				
12	2012 Dead in 2015 50 ok				
13	2012 Alive in 2015 103 ok				
14	2013 Dead in 2015 50 ok				
15	2013 Alive in 2015 103 ok				
16	2014 Dead in 2015 50 ok				
17	2014 Alive in 2015 103 ok				
18	2015 Dead in 2015 50 ok				
19	2015 Alive in 2015 103 ok				
20	ACRO validated output				
21	-----				
22	survivor				
23	year Dead in 2015 Alive in 2015				
24	-----				
25	2010 128 247				
26	2011 128 247				
27	2012 128 247				

Figure 9 Example of weighted output

As can be seen in Figure 9, the underlying counts are for 50 or 103 observations in each year. These are checked against the SDC rules. However, the full Stata output, from line 20 onwards, has the weighted results as requested.

For estimation results, the SDC-summary takes a different form:

	A	B	C	D	E	F	G	H	I	J
1	Rules applied: Default									
2	regress inc_activity inc_grants inc_donations total_costs									
3	Outcome: pass									
4	Back to description page									
5		inc_grants	inc_donat	total_cost	_cons					
6	b	-0.89	-0.67	0.83	#####					
7	se	0.02	0.02	0.01	#####					
8	t	-36.13	-40.91	78.94	0.75					
9	pvalue	0.00	0.00	0.00	0.45					
10	ll	-0.93	-0.70	0.81	#####					
11	ul	-0.84	-0.63	0.85	#####					
12	df	807.00	807.00	807.00	807.00					
13	crit	1.96	1.96	1.96	1.96					
14	eform	0.00	0.00	0.00	0.00					
15										
16										
17										
18										
19	ACRO validated output									
20	Source	SS	df	MS	Number of obs	=	811			
21					F(3, 807)	=	2276.25			
22	Model	1.3357e+18	3	4.4524e+17	Prob > F	=	0.0000			
23	Residual	1.5785e+17	807	1.9560e+14	R-squared	=	0.8943			
24					Adj R-squared	=	0.8939			
25	Total	1.4936e+18	810	1.8439e+15	Root MSE	=	1.40107			

Figure 10 Example estimation output

The SDC-summary shows a table of coefficients, standard errors, t- and z-values etc. for each variable. The usual Stata output is presented below. Although the number of rows in this block varies with the estimation type, it always starts on the same line. This means researchers can extract relevant information for re-presentation reliably and more easily than from the standard Stata output.

For graphs, there is no SDC-summary, as each graph has to be manually reviewed. Instead, a sheet for a graph output contains simply the command line and the graph itself:

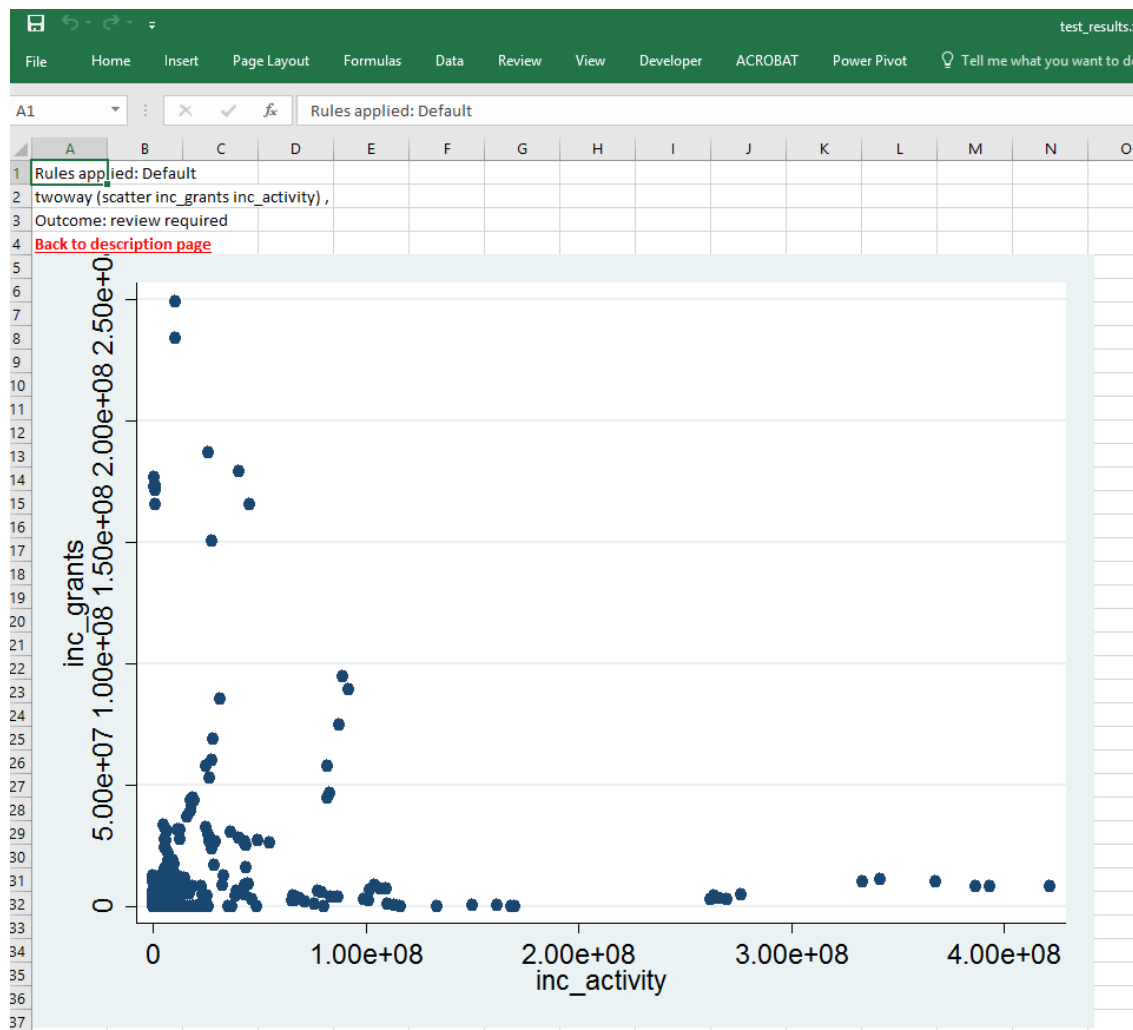


Figure 11 Example graph output

4. Using ACRO – steps for the researcher to take

ACRO has four stages for the researcher to carry out

- Preparation: Set up your ado-path to find all the ACRO files.
- Set-up: run the command **safe_setup**. This defines the output workbook, preferences for suppression, and the dataset being used (some datasets have specific rules). Default values are used if the researcher does not specify them.
- Sending outputs for checking: prefix each command that produces an output with **safe**, to have it cleared for release. Options for exceptions, suppression and sheet name are already set in the set-up stage. Apart from these options, there is no difference between the STATA command and the **safe** command. If the command is not recognised, then the code prints an error message and ignores the command. This stage can be run repeatedly: outputs with the same name are overwritten.
- Finalising: run the command **safe_finalise**. This writes the index file for the workbook, allowing the output checker to easily see which output needs review.

The ACRO commands can be called interactively in a command session, or in a program. At any point, the researcher can examine the workbook being created by ACRO and review his or her

activities. Note however that you **must close the workbook when running ACRO**; this is because Excel won't let another program make changes to an open workbook.

The examples below are generated by the programs **test_file_full.do** and **test_file_small.do**, which are found in the folder with all the ACRO ado-files. Researchers should copy these file to their own workspace, edit the file path for their own environment, and experiment with the settings. A sample data set **test_data.dta** can be downloaded from the ado-file folder¹.

We now consider these in more detail. See the test files for practical examples of using ACRO.

4.1 Findings files and checking the SDC rules

ACRO consists of a large number of Stata ado files. These may already be included in your Stata path. If not, you need to add the ACRO folder to your ado-path. For example, if the ACRO files are held in the folder H:\documentation\ACRO, then you need to have this at the start of your Stata code:

```
adopath + "X:\documentation\ACRO"
```

This will set the correct 'adopath' (see test file for an example).

To test whether the system manager has set up the ado-path for you (or to see if you have set it up correctly), once you have run **safe_setup** (see below) enter

```
safe_show_SDC_parameters
```

This shows the SDC parameters that will be applied for a particular dataset, or the default values if no dataset is specified. If this command works, the ado-path has been set up correctly.

4.2 Set-up

To begin the process, call the program **safe_setup** with four parameters:

```
safe_setup working_folder default_results_file [suppress]
```

The order of the parameters matters. The parameters are

- **working_folder**: where the results file and temporary files are to be stored
- **SDC rules to apply**: SDC rules can vary between datasets, so you may be told by your system manager to use a particular value such as 'ESS' or 'social'; 'default' should be used unless you have been told to use a specific name
- **results_file**: the name of the output workbook (without the "xlsx" or "xls" extension)
- **suppress (optional)**: whether the default position is to suppress unacceptable outputs, where feasible; if this option is not selected, then the default is not to suppress outputs.

For example:

```
safe_setup "C:\temp\output checking" CIS test_results  
suppress
```

This tells ACRO that temporary results and the final output are to be stored in **C:\temp\output_checking**, that the SDC rules relevant to the CIS should be applied, that the

¹ This is the dataset used in Green, E., Ritchie, F., Bradley, P., & Parry, G. (2021) "Financial resilience, income dependence and organisational survival in UK charities". *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, <https://doi.org/10.1007/s11266-020-00311-9>. Available from <https://uwe-repository.worktribe.com/output/6975797>

resulting workbook will be **test_results.xlsm**, and that the default is to apply suppression where relevant.

Note that suppression is not feasible for all output. For example, it is feasible to suppress table cells, but for regression results either the regression as a whole passes or it does not; therefore, suppression of elements is not relevant.

If **safe_setup** has not been run before any outputs are generated, then Stata will ignore the **safe** requests and print a message

```
*** Safe setup not yet run ***
```

4.3 Sending outputs for checking

Assume that the command to be run is a two-way tabulation (year, survival status) of actual and expected frequencies with chi-square test statistics automatically calculated. This would normally be written as

```
tab year survivor, chi2 expected
```

Running ACRO consists of prefixing the command with **safe** and postfixing with up to three options specific to ACRO:

```
safe tab year survivor, chi2 expected
[output_sheet("name")] [exception("reason")]
[suppress|nosuppress]
```

The options are:

- `output_sheet("name")` is the name of the spreadsheet in the Excel file to be created. If the sheet already exists, it will be overwritten. If this option is omitted, then the sheet will be automatically named "output_1", "output_2" etc.
- `exception("reason")` means that the researcher has requested that, if this command fails SDC checks, then it should be considered as an exception to those rules for the reasons stated in "reason". The reasons should be specific to that output (eg "min and max age are structural" rather than general "I don't think there are any problems here"). If the output passes the SDC checks, then this is not used
- `(no)suppress` allows the researcher to override the default suppression option set in the set-up phase

So to run the above command with

- results going to the sheet "yr-surv counts"
- default suppression behaviour, and
- making a case for low sensitivity of the results in the case of a failure

The original (not checked) command would be:

```
tab year survivor, chi2 expected
```

But the command to produce a checked output would be

```
safe tab year survivor, chi2 expected output_sheet("yr-
surv counts") exception("I think these are not
sensitive")
```

The order of the options does not matter.

The options “exception” and “suppress” are only relevant to some commands, where some cells might be acceptable and other cells might fail. Removing unacceptable cells and retaining others is a valuable outcome.

In the case of estimation, exceptions and suppression are not relevant, and so only the (optional) output sheet is required. Thus

```
probit survivor inc_activity inc_grants inc_donations
      total_costs
```

can be rendered as

```
safe probit survivor inc_activity inc_grants
      inc_donations total_costs ,
      output_sheet("reg_result")
```

or just

```
safe probit survivor inc_activity inc_grants
      inc_donations total_costs
```

The SDC test for estimates is whether there are sufficient residual degrees of freedom in the model². This is a yes/no question for the model as a whole: partial suppression is not relevant. There are no meaningful exceptions.

For graphs, no SDC checks are carried out as currently these need manual review. However, the functionality to load graphs into the spreadsheet is included in ACRO as this allows a wider variety of output to be packaged into the same workbook.

4.4 Finalising

Once the output have been requested, the researcher **must** call

```
safe_finalise
```

This writes the index for the Excel spreadsheet and deletes temporary files. There are no parameters. **If this is not called**, the index for the workbook will not be written, and the workbook will be rejected by the output checker.

5. Hints

safe can be called interactively from the command line, or written into programs. The former works well for testing, but the latter gives much more flexibility. Macros can be used to allow programmers to switch **safe** on and off easily – see the test files for examples.

The SDC checks on tables can slow programs down, particularly with large datasets and with many sub-tables. If a program is using ‘by’ and is running slowly, it might be more efficient to reprogram as a series of ‘if’ statements.

² For the rationale for this rule, and all of the other SDC rules, see Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010), *Guidelines for the checking of output based on microdata research*, Final report of ESSnet sub-group on output SDC. <http://eprints.uwe.ac.uk/22487/>

Safe needs to be called 'noisily' to run; this is because it needs to generate text in a log file. Coders wanting to run code in quiet mode can do so, but may need to switch into noisy mode for 'safe' commands:

```
noisily {  
    safe probit survivor inc_activity inc_grants  
        inc_donations total_costs  
}
```

An indication that the code needs to run noisily is if the code stops with error "**v1 not found**"; this indicates that Stata has written no output because it is running in quiet mode. Unfortunately, this can't be overwritten within **safe** because of the way Stata treats code files as part of a hierarchy of programs.