ACRO
Automated checking of research output

The 2021 joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality
1-3 December 2021.
Introduction

- Pilot project commissioned by Eurostat in 2019
- Delivered in 2021 by UWE* – Bristol
- Proof-of-concept software based on STATA®
- Designed as a tool capable to reduce the workload of output checking, traditionally made manually by output checking officers
- Published as Statistical Working Paper “Automatic Checking of Research Outputs (ACRO): a tool for dynamic disclosure checks”
- Source code published on https://github.com/eurostat/ACRO

* F.Ritchie, J.Smith and E.Green
Functional features

1. Implementation of automatic checks on:
   • Tabulations
   • Common estimators
   • Medians
   • Maxima/Minima (simple ban)
2. Automatic primary suppression
3. Requests of exceptions
4. Report preparation (Excel)
## Statistical Disclosure rules applied

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
<th>Applies to</th>
<th>Commands</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Threshold</strong></td>
<td>Minimum number of observations underlying a statistic</td>
<td>All linear statistics: frequencies, mean, median, sums etc</td>
<td>table tabulate</td>
</tr>
<tr>
<td><strong>N-K dominance</strong></td>
<td>The N largest observations should not count for more than K% of the total</td>
<td>As for threshold, but doesn’t apply to frequencies</td>
<td>table tabulate</td>
</tr>
<tr>
<td><strong>P-ratio</strong></td>
<td>It should not be possible to estimate another observation within p% of its value</td>
<td>As for N-K rule</td>
<td>table tabulate</td>
</tr>
<tr>
<td><strong>Table rule</strong></td>
<td>SDC rules are applied to each table cell independently; any cell can pass or fail</td>
<td>Applies to all tables</td>
<td>table tabulate</td>
</tr>
<tr>
<td><strong>Maximum &amp; minimum</strong></td>
<td>Not allowed</td>
<td>Any magnitude</td>
<td>table tabulate</td>
</tr>
<tr>
<td><strong>Degrees of freedom</strong></td>
<td>Analytical outputs must have at least K degrees of freedom</td>
<td>Analytical results, including estimation and testing</td>
<td>regress, xtreg logit, probit test, ttest</td>
</tr>
</tbody>
</table>
Features not implemented

1. Automatic checks on graphs (all set to review)
2. Percentiles (except median)
3. Secondary (within table) suppression
Practical use

A researcher:
• starts the STATA® platform, imports the data
• runs a ‘setup’ script
• uses the keyword ‘safe’ to prefix the commands
• uses ‘finalise’ to prepare the spreadsheet for review by the output checker.
### Sample ACRO output

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>![x]</td>
<td>![✓]</td>
<td>![fx]</td>
<td>=HYPERLINK(&quot;[test_results.xlsx]!A1&quot;,&quot;activity&quot;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Sheet</strong></td>
<td><strong>Automatic check</strong></td>
<td><strong>Final decision</strong></td>
<td><strong>Description safe/unsafe</strong></td>
<td><strong>Reason for automatic decision</strong></td>
<td><strong>Exception request</strong></td>
</tr>
<tr>
<td>activity</td>
<td>ok</td>
<td>ok</td>
<td>unsafe statistic: table</td>
<td>pass</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>graph test</td>
<td>review</td>
<td></td>
<td>graph: twoway</td>
<td>review required</td>
<td></td>
<td></td>
</tr>
<tr>
<td>max act</td>
<td>ok</td>
<td>ok</td>
<td>unsafe statistic: table</td>
<td>fail; suppression</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 1</td>
<td>ok</td>
<td>ok</td>
<td>unsafe statistic: tabulate</td>
<td>pass</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 2 A</td>
<td>fail</td>
<td>fail</td>
<td>unsafe statistic: tabulate</td>
<td>fail</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 2 B</td>
<td>ok</td>
<td>ok</td>
<td>unsafe statistic: tabulate</td>
<td>pass</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 2 C</td>
<td>ok</td>
<td>ok</td>
<td>unsafe statistic: tabulate</td>
<td>pass</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 2 D</td>
<td>fail</td>
<td>fail</td>
<td>unsafe statistic: tabulate</td>
<td>fail</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 3</td>
<td>fail</td>
<td>fail</td>
<td>unsafe statistic: table</td>
<td>fail</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>output 4</td>
<td>review</td>
<td></td>
<td>unsafe statistic: table</td>
<td>fail; exception request</td>
<td></td>
<td>I'm a professor</td>
</tr>
<tr>
<td>output 5</td>
<td>ok</td>
<td>ok</td>
<td>safe statistic: regress</td>
<td>pass</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td>small act A</td>
<td>review</td>
<td></td>
<td>unsafe statistic: table</td>
<td>fail; exception request</td>
<td></td>
<td>It's not feasible to identify the charities from this information</td>
</tr>
<tr>
<td>small act B</td>
<td>review</td>
<td></td>
<td>unsafe statistic: table</td>
<td>fail; exception request</td>
<td></td>
<td>It's not feasible to identify the charities from this information</td>
</tr>
<tr>
<td>small act C</td>
<td>review</td>
<td></td>
<td>unsafe statistic: table</td>
<td>fail; exception request</td>
<td></td>
<td>It's not feasible to identify the charities from this information</td>
</tr>
</tbody>
</table>
Advantages

• Overall good efficiency
• Minimal setup required
• Low training overhead
• Readability of output results (MS Excel® files)
Limitations

- Only one language available (Stata scripts)
- Narrow range of implemented functions
Initial review

Initial reviewers* find that ACRO:
• has good potential to reduce output checking workload
• is not perfectly intuitive as expected, though it can be learnt with a minimal effort
• implementing a wider range of functions would substantially improve the utility of the tool
• porting the tool to a wider set of languages (R/SAS) is necessary

(*) SDC experts and managers of RDCs
Next steps

• **We welcome feedback in order to guide next developments:**
  - Feedback by email: estat-confidentiality@ec.europa.eu
  - Feedback by Git: fork the project, work your branch, then pull request

• **Further developments to address limitations and improve general functionalities of the tool and its buy-in by RDC managers and users.**
References


ACRO
Automated checking of research output

Thank you for your attention!

The 2021 joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality 1-3 December 2021.