SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt
Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

ABSTRACT

This paper introduces a macro (**case_control_check**) that tracks the number of records/cases/controls in a dataset and how those change over the course of a project.

What are the biggest issues programmer faces when writing scripts?

Are there any dropped or added records, or are there any duplicates, especially after linking multiple datasets. The SAS log is available, but does not provide all the information. A programmer has to write additional code after each data or proc step. This will be a cumbersome task for bigger projects.

The purpose of this macro is to report on these issues with minimal amount of extra work. This macro was developed for epidemiological (case-control) studies but has been applied elsewhere. This paper explains macro parameters followed by an example to better understand the functionality of this macro.

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt
Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

Intro

This macro (case_control_check) tracks the number of records/cases/controls change after each data/proc step as well as check for duplicates. It automatically takes the last dataset from the “syslast” macro variable and uses it in conjunction with user-provided info, like primary key and what defines a record, case or control, (e.g. is_case = 1 defines a case; Id and date could be the primary keys). It creates output in both dataset and text file format displaying the dataset name, the number of records/cases/controls in the dataset, the change in the number of records from last macro call, the number of duplicates and reason for that change (if provided by the user).

Macro parameters:

- **macro_num**= Start with 1 and provide number in the increment of 1 for each macro call.
- **case_definition**= Variables which define a case i.e is_case = 1. May use %str.
- **case_change_reason**= If you know that number of cases will change from last step, an optional reason for future reference. May use %str.
- **case_primary_key**= Variables which define a unique case i.e Id index_date. It will be used in conjunction with case_definition parameter.
- **control_definition**= Variables which define a control i.e is_case = 0 (Optional). May use %str.
- **control_change_reason**= If you know that number of controls will change from last step, an optional reason for future reference. May use %str.
- **control_primary_key**= Variables which define a unique control i.e Id index_date. It will be used in conjunction with control_definition parameter (Optional).
- **filename_prefix**=Name of the output file, which should be same in each call of the macro within the same project. A datetime timestamp will be suffixed automatically. Use no special characters, spaces or hyphen, "_" is preferred.
- **file_location**=Location of the output file and should be same in each call of the macro within same project.

VACCINE & DRUG Evaluation Centre

University of Manitoba

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt

Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

Example code

```
%let file_loc = C:\data;
/*Create population dataset*/
data population_dtset;
    set sashelp.baseball;
    where nAtBat >= 200;
run;
/* First call to the macro*/
%case_control_check(macro_num=1,
    case_definition=%str(not missing(name)),
    case_change_reason=Initial run,
    case_primary_key=name,
    filename_prefix=baseball,
    file_location=&file_loc.);
/*Create case dataset from population dataset*/
data case_dtset;
    set population_dtset;
    where nAtBat >= 500
      and YrMajor >= 5;
    is_case = 1; /*Create new variable is_case and assign value 1 to it*/
run;
/*Second call to the macro*/
%case_control_check(macro_num=2,
    case_definition=%str(is_case=1),
    case_change_reason=Case selection,
    case_primary_key=name,
    filename_prefix=baseball,
    file_location=&file_loc.);
/*Create controls dataset who are in same league, division and plays at the
    same position but times at bat (natbat) is less than 500*/
proc sql;
    create table control_dtset as
    select a.*,
           0 as is_case
    from population_dtset a
        , case_dtset b
    where a.league   = b.league
      and a.division = b.division
      and a.Position= b.Position
      and a.nAtBat < 500
      and a.name not= b.name
    ;
quit;
```

```
/*Create cohort*/
data cohort_dtset_1;
    set case_dtset
        control_dtset;
run;
/*Third call to the macro*/
%case_control_check(macro_num=3,
    case_definition=%str(is_case=1),
    case_primary_key=name,
    control_definition=%str(is_case = 0),
    control_change_reason=Controls selection,
    control_primary_key=name,
    filename_prefix=baseball,
    file_location=&file_loc.);
/*Remove anyone from cohort with missing salary*/
data cohort_dtset;
    set cohort_dtset_1;
    where not missing(salary);
run;
/*Fourth call to the macro*/
%case_control_check(macro_num=4,
    case_definition=%str(is_case=1),
    case_change_reason=Remove missing salary records,
    case_primary_key=name,
    control_definition=%str(is_case = 0),
    control_change_reason=Remove missing salary records,
    control_primary_key=name,
    filename_prefix=baseball,
    file_location=&file_loc.);
```

Input dataset (sashelp.baseball)

Name	Team	nAtBat	YrMajor	League	Division	Position	Salary
Allanson, Andy	Cleveland	293	1	American	East	C	.
Ashby, Alan	Houston	315	14	National	West	C	475
Davis, Alan	Seattle	479	3	American	West	1B	480
Dawson, Andre	Montreal	496	11	National	East	RF	500
Galarraga, And...	Montreal	321	2	National	East	1B	91.5
Griffin, Alfredo	Oakland	594	11	American	West	SS	750
Salazar, Argenis	Kansas City	298	3	American	West	SS	100
Thomas, Andres	Atlanta	323	2	National	West	SS	75
Thornton, Andre	Cleveland	401	13	American	East	DH	1100
Trammell, Alan	Detroit	574	10	American	East	SS	517.143
Trevino, Alex	Los Angeles	202	9	National	West	C	512.5
Van Slyke, Andy	St Louis	418	4	National	East	RF	550

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt
Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

VACCINE & DRUG Evaluation Centre

University of Manitoba

Output

	dataset_name	cases	changed_cases	Reason_cases	duplicates_in_cases	controls	changed_controls	Reason_controls	duplicates_in_controls
1	WORK.POPULATION_DTSET	293	0	Initial run	0	0	0		0
2	WORK.CASE_DTSET	65	-228	Case selection	0	0	0		0
3	WORK.COHORT_DTSET_1	65	0		0	229	229	Controls selection	114
4	WORK.COHORT_DTSET	59	-6	Remove missing salary records	0	180	-49	Remove missing salary records	90

There are duplicate records which means either we didn't defined the correct primary key or there is a bug in the code.

```
/*Create controls dataset who are in same league, division and plays at the
    same position but times at bat (natbat) is less than 500*/
proc sql;
    create table control_dtset as
    select distinct a.*,
           0 as is_case
    from population_dtset a
        , case_dtset b
    where a.league   = b.league
      and a.division = b.division
      and a.Position= b.Position
      and a.nAtBat < 500
      and a.name not= b.name
    ;
quit;
```

After correcting the bug in the code (used distinct when creating control_dtset), the duplicates are gone.

Output after rerunning the code

	dataset_name	cases	changed_cases	Reason_cases	duplicates_in_cases	controls	changed_controls	Reason_controls	duplicates_in_controls
1	WORK.POPULATION_DTSET	293	0	Initial run	0	0	0	"	0
2	WORK.CASE_DTSET	65	-228	Case selection	0	0	0	"	0
3	WORK.COHORT_DTSET_1	65	0	"	0	115	115	Controls selection	0
4	WORK.COHORT_DTSET	59	-6	Remove missing salary records	0	90	-25	Remove missing salary records	0

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt
Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

Output SAS dataset and text file

The output SAS data set or text file contains the following variables:

- **Dataset** = Name of the dataset
- **No. of cases** = Number of cases based on case_definition parameter
- **Cases changed** = Change in number of cases from previous step
- **Cases change reason** = Reason of change if provided in case_change_reason parameter
- **Cases duplicates** = Duplicates in cases using case_primary_key and case_definition parameters
- **No. of controls** = Number of controls based on control_definition parameter
- **Controls changed** = Change in number of controls from previous step
- **Controls change reason** = Reason of change if provided in control_change_reason parameter
- **Controls duplicates** = Duplicates in controls using control_primary_key and control_definition parameters

The output SAS dataset created is "**tracker_ds**".
The text file output is suffixed automatically with a datetime timestamp. The macro will automatically use the most recent text file if required.

Text file output

baseball_18NOV19112043.txt - Notepad

File Edit Format View Help

Dataset	No. of cases	Cases changed	Cases change reason	Cases dups	No. of controls	Controls changed	Controls change reason	Controls dups
WORK.POPULATION_DTSET	293	0	Initial run	0	0	0	"	0
WORK.CASE_DTSET	65	-228	Case selection	0	0	0	"	0
WORK.COHORT_DTSET_1	65	0	"	0	115	115	Controls selection	0
WORK.COHORT_DTSET	59	-6	Remove missing salary records	0	90	-25	Remove missing salary records	0

4772 - A program to keep track of the number of records in a project

Gurpreet Pabla, Christiaan Righolt
Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba

Conclusion

- Automatic tracking of population
- Leaves audit trail
- Comparison with earlier results
  - New code
  - New/extended data
  - Different project
- Does not require access to code
- Understandable by non-programmer/non-analysts

Contact Information

Please contact below for the macro:

- Gurpreet.Pabla@umanitoba.ca
- Christiaan.Righolt@umanitoba.ca

MARCH 29 – APRIL 1 | WASHINGTON, DC | #SASGF