



SAS - By Group Processing

umanitoba.ca/centres/mchp



UNIVERSITY
OF MANITOBA





Winnipeg SAS users Group

SAS – By Group Processing

Are you First or Last In Line
Charles Burchill

Manitoba Centre for Health Policy,
University of Manitoba



What is By Group Processing

- Records are grouped into similar sets based on one or more variables in a dataset
- Uses
 - Combining (merge/join) two or more sets of data
 - Summarized information based on by groups.



Sample Data: ATC_DINS

ATC	DIN	DATE	English Name
A07BC04	02229951	20000517	KAOPECTATE
A07BC04	02229951	20040416	KAOPECTATE
A07BC04	02229951	20060920	KAOPECTATE
A07BC04	02229951	20080814	KAOPECTATE
A07BC04	02229953	20000517	KAOPECTATE
A07BC30	00238651	19990413	KAOLIN MIXTURE W PTN
A07BC30	00411949	19990412	DIAREX
A07BC51	00373656	20000517	WATKINS SETTELZ
A07BC51	00373656	20060720	WATKINS SETTELZ



Pre-Processing

- Sort
 - Use proc sort to order the data appropriately.
- Index
 - If an index has been added to the data for the variables of interest it does not need to be sorted.
 - Indexes are added using
 - Proc Datasets with modify; index,
 - Proc SQL using create index on command
 - Data step (index=()) option



Syntax

BY <DESCENDING> variable(s)
<NOTSORTED> <GROUPFORMAT>;



Data Step Processing (First/Last)

Temporary Variables

- Automatically created when BY statement used
- Each level of by variable, nested in order
- First.xxxx, Last.xxxx

```
Data atc_dins ;  
    set atc_dins ;  
    by atc din ; * creates first.atc, last.atc, first.din last.din ;  
Run;
```



Temporary Variables

ATC	DIN	FIRST.ATC	LAST.ATC	FIRST.DIN	LAST.DIN
A07BC04	02229951	1	0	1	0
A07BC04	02229951	0	0	0	0
A07BC04	02229951	0	0	0	0
A07BC04	02229951	0	0	0	1
A07BC04	02229953	0	1	1	1
A07BC30	00238651	1	0	1	1
A07BC30	00411949	0	1	1	1
A07BC51	00373656	1	0	1	0
A07BC51	00373656	0	1	0	1



Selecting (Data Step Processing)

- If FIRST.ATC then ... ;
Or If LAST.DIN then ... ;

```
Proc sort data=atc_dins ;  
  by atc din date ;
```

```
Run;
```

```
Data atc_dins ;
```

```
  set atc_dins ;
```

```
  by atc din ; * creates first.atc, last.atc, first.din last.din ;
```

```
  if last.din then output ; ** output 1 record/din based on  
  most recent but may just create a variable to flag in data;
```

```
Run;
```



Data step processing

The Data step loop – a quick review

- 1. Program Data Vector Created

Contains variables from input data, created variables, automatic variables (`_n_`, `_error_`, `first.x`, `last.x`, etc...)

All variable values are missing (“”, or .)

- 2. Values from input data loaded to PDV, derived or calculated

- 3. PDV Written to output data on last statement or output.

- 4. Back to 1. 



Use of RETAIN (Data Step Processing)

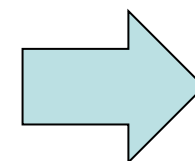
- RETAIN is used to hold the value of a variable from one data step iteration to the next.

```
data atc_din_out ;  
  set atc_din ;  
  by atc din ;  
  retain first_date ;  
  if first.din then first_date=start_date ;  
  if last.din then output ;  
  format first_date yymmddn8. ;  
  rename start_date=last_date ;  
  
run;
```



ATC_DIN DATA

ATC	DIN	DATE	FIRST. ATC	LAST. ATC	FIRST. DIN	LAST. DIN	First_date
A07BC04	02229951	20000517	1	0	1	0	20000517
A07BC04	02229951	20040416	0	0	0	0	20000517
A07BC04	02229951	20060920	0	0	0	0	20000517
A07BC04	02229951	20080814	0	0	0	1	20000517
A07BC04	02229953	20000517	0	1	1	1	20000517
A07BC30	00238651	19990413	1	0	1	1	19990413
A07BC30	00411949	19990412	0	1	1	1	19990412
A07BC51	00373656	20000517	1	0	1	0	20000517
A07BC51	00373656	20060720	0	1	0	1	20000517





Use of RETAIN – summary/conditional

```
data atc_din_out ;
  set atc_din ;
  by atc din ;
  retain counter multi_rec first_date;
  if first.din then do ;
    counter=0 ;
    first_date = start_date ;
  end ;

  counter = counter + 1 ;
  if 2<=counter <10 then Multi_rec = 1 ;
  else if 10<=counter < 100 then Multi_rec= 10 ;
  else counter >= 100 then Multi_rec= 100 ;
  Time_since_start = start_date - first_date ;

run;
```



BY Options

BY <DESCENDING> variable(s) <NOTSORTED> <GROUPFORMAT>;

- **NOTSORTED**

- Data does not need to be sorted, each change in BY variables generates new “BY” group
- The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Not available with Merge, Update, or Set with more than one dataset
- Used on By statements Data step and procedures

- **GROUPFORMAT**

- First and Last assigned by the formatted values
- Only available in data step processing. It is the same as BY-group processing with formatted values in SAS procedures.



Use in Procedures

- **BY**
 - Run analysis by each level found in the by variable values
 - Requires data to be sorted but more efficient than class
 - Most procedures
- **CLASS**
 - Some procedures run separate analysis by each level in the class variable (means, univariate, tabulate, summary, timeplot, ttest)
 - Additional options
 - Data does not need to be sorted by the BY variable
 - Other procedures (e.g. GENMOD) specify categorical variables



Use in Procedures

```
Proc Means data=hospital ;  
    by region sex ;  
    var los ;  
Run;
```



SQL Processing (summary functions)

```
PROC SQL ;  
    SELECT DISTINCT ATC, DIN,  
        MIN(START_DATE) AS FIRST_DATE FORMAT=YYMMDD8.,  
        MAX(START_DATE) AS LAST_DATE FORMAT=YYMMDD8.  
    FROM ACT_DIN  
    GROUP BY ATC, DIN ;  
QUIT ;
```



Some Sources of BY information

- SAS documentation: BY-Group Processing in the DATA Step
 - <http://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm#a001283274.htm>
- SGF 2007: The Power of the BY Statement
 - <http://www2.sas.com/proceedings/forum2007/222-2007.pdf>
- SUGI 29 2004: Creating and Exploiting SAS Indexes
 - <http://www2.sas.com/proceedings/sugi29/123-29.pdf>
- SUIG 27 2002: Longitudinal Data Techniques: Looking Across Observations
 - <http://www2.sas.com/proceedings/sugi27/p015-27.pdf>
- NESUG 2007: Summing with SAS
 - <http://www.nesug.org/proceedings/nesug07/ff/ff08.pdf>
- TU 2009: Retaining, Lagging, Leading, and Interleaving Data
 - http://changchung.com/download/retainLagLeadInterleave_draft.pdf



Thank You!

