PharmaSUG 2012 - Paper PO02

Does NODUPKEY Select the First Record in a By Group?

David Franklin, TheProgrammersCabin.com, Litchfield, NH

ABSTRACT

Does the NODUPKEY option in the SORT procedure always select the first observation in a group of variables? There are some who say it does and some that say it does not.

This paper looks into this question, with examples, and shows that the NODUPKEY has really no effect on whether the first observation in a group of data gets selected, but does find that there are two other options that effect whether this happens and presents the effects of these options. Because these two options are rarely specified in programming and the default values shipped with SAS are almost always used, if these values are changed, unexpected results will happen.

INTRODUCTION

For some, the NODUPKEY option in the SORT procedure always selects the first observation in a group of variables. But does it? This paper delves into this question and finds that the NODUPKEY actually has no real effect, but two other options do!

FIRST, SOME DATA

For the purposes of this paper, the following data is used to produce the output:

```
data ae0;
   input ptnum $6. aeseq event $20.;
cards;
001002 1 HEADACHE
001001 1 FEVER
001001 2 HEADACHE
001003 1 NAUSEA
001003 4 DIARRHEOA
001003 2 VOMITING
001004 1 DIARRHEOA
001001 3 DIARRHEOA
001002 2 DIARRHEOA
001004 2 HEADACHE
001003 3 FEVER
;
run;
```

OUR FIRST LOOK USING NODUPKEY

First, a simple basic SORT procedure call that outputs the unique PTNUM values:

```
proc sort data=ae0 out=ae1 nodupkey;
   by ptnum;
proc print data=ae1;
   title1 "Output Using NODUPKEY";
run;
```

The output is given below:

```
                     Output Using NODUPKEY

              Obs    ptnum     aeseq    event

               1     001001      1      FEVER
               2     001002      1      HEADACHE
               3     001003      1      NAUSEA
               4     001004      1      DIARRHEOA
```

Output 1. Output from PROC SORT with NODUPKEY Option

Does this produce the first observation of each PTNUM value? It depends on whether you expect the first observation in each PTNUM group to have a AESEQ value of 1 or not – to quote the SAS documentation, “or if the observations with identical BY variable values are to retain the same relative positions in the output data set as in the input data set”. Now welcome two options that make this happen!

THE TWINS …

Welcome the EQUALS|NOEQUALS option in the SORT procedure, and the global option SORTEQUAL|NOSORTEQUALS – it is these two options that effect whether the first observation in a group is selected or not.

The documentation for the SORTEQUAL|NOSORTEQUALS is given below:

SORTEQUALS

> specifies that observations with identical BY variable values are to retain the same relative positions in the output data set as in the input data set.

NOSORTEQUALS

> specifies that no resources be used to control the order of observations with identical BY variable values in the output data set.

Their counterparts in the SORT procedure are EQUALS for SORTEQUALS and NOEQUALS for the NOSORTEQUALS. The default value when SAS is shipped is SORTEQUALS.

It is interesting to note that SAS suggests that in order to save resources, use NOSORTEQUALS when you do not need to maintain a specific order of observations with identical BY variable values.

… AND OH WHAT TROUBLE THEY CAN MAKE

Now lets look at what these options can do – first with the EQUALS option used:

```
proc sort data=ae0 out=ae1 nodupkey equals;
   by ptnum;
proc print data=ae1;
   title1 "Output Using NODUPKEY and EQUALS Options";
run;
```

Will result in the following output:

```
           Output Using NODUPKEY and EQUALS Options

              Obs    ptnum     aeseq    event

               1     001001      1      FEVER
               2     001002      1      HEADACHE
               3     001003      1      NAUSEA
               4     001004      1      DIARRHEOA
```

Output 2. Output from PROC SORT with NODUPKEY and EQUALS Option

This is the output we got previously.

Now lets look at if the EQUALS option was changed to the NOEQUALS option:

```
proc sort data=ae0 out=ae1 nodupkey noequals;
   by ptnum;
proc print data=ae1;
   title1 "Output Using NODUPKEY and NOEQUALS Options";
run;
```

Will result in the following output:

```
             Output Using NODUPKEY and NOEQUALS Options

               Obs     ptnum     aeseq     event

                1      001001      3       DIARRHEOA
                2      001002      1       HEADACHE
                3      001003      3       FEVER
                4      001004      1       DIARRHEOA
```

Output 3. Output from PROC SORT with NODUPKEY and NOEQUALS Option

The output now is not the first record in the output sequence, nor is it is the first record found for each group, i.e. PTNUM value.

Similar results are attained using the global SORTEQUALS|NOSORTEQUALS options.

One advantage of using the NOEQUALS or NOSORTEQUALS option is that the sort procedure saves resources, which can be valuable when dealing with large datasets.

CONCLUSION

In answer to the original question "Does the NODUPKEY option in the SORT procedure always select the first observation in a group of variables?" then answer is "Not Really" – the two options that do effect this are the EQUALS|NOEQUALS options in the SORT procedure, and the global options SORTEQUALS|NOSORTEQUALS. As this paper has shown, altering these options can effect the output, so if your program relies on the first observation in dataset being selected using the NODUPKEY option, specify one of these two options where appropriate.

It is always a good idea to remember that inside SAS there are a large number of SAS options that can affect the output that is produced, whether is be from a procedure or datastep -- if your program produces unexpected results, after you have gone though looking at the SAS LOG and exhausted other remedies, take a look at the SAS options that are applied.

REFERENCES

SAS Institute Inc. 2006. Base SAS® 9.1.3 Procedures Guide, Second Edition, Volumes 1, 2, 3, and 4. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: David Franklin
Enterprise: TheProgrammersCabin.com
Address: 16 Roberts Rd
City, State ZIP: Litchfield, NH 03052
Work Phone: 603-275-6809
Email: dfranklin@TheProgrammersCabin.com
Web: TheProgrammersCabin.com