Simulate PRELOADFMT Option in PROC FREQ

Ajay Gupta, PPD, Morrisville, NC

ABSTRACT

In Pharmaceuticals/CRO industries, table programing is often started when only partial data is available and it is common that we need to summarize data based on all possible combinations of values. However, if there is no data collected for all levels of these values in the data set, by default, existing SAS procedures are not able to display them in a summary table. A viable solution is to add the PRELOADFMT option in these procedures. PROC FREQ is commonly used procedure to summarize data and calculate statistics, but the PRELOADFMT option is not supported. However using the existing SAS procedures (e.g. PROC FORMAT, PROC DATASETS, and PROC FREQ), we can create a SAS macro which can perform the above mentioned task easily without much programming work.

INTRODUCTION

Table programing is often started when only partial data is available and it is common that we need to summarize data based on all possible combinations of values. However, if there is no data collected for all levels of these values in the data set, by default, existing SAS procedures are not able to display them in a summary table. A viable solution is to add the PRELOADFMT option in these procedures. This option specifies that all formats are preloaded with all possible levels for all the variables of interest. But some procedures like PROC FREQ, which are widely used to summarize data and calculate statistics, do not have the PRELOADFMT option.

In order to simulate the PRELOADFMT option in a PROC FREQ, for example, a viable solution would be to pre-process the data with all plausible combinations of values and later execute the PROC FREQ using the SPARSE option in the TABLE statement.

HANDLING MISSING VALUES USING PROC FREQ

This paper will identify the issue by executing PROC FREQ on demographic data with missing values. Later, the paper will discuss above mentioned solution step by step.

CREATE DEMOGRAHIC DATA AND FORMATS

For demonstration purpose, we will use demographic data set with following scenarios:
- The data possibly can have three treatment groups. In this example treatment 3 is not present in the data set
- The data possibly can have both male and female subjects. In this example there are no male subjects in Treatment 2.
- There will be one associated format for gender and treatment each.

```
/*Create format for treatment group and gender*/

PROC FORMAT;
      VALUE $trt
            "1"="Treatment 1"
            "2"="Treatment 2"
            "3"="Treatment 3"
      ;
      VALUE $sex
            "1"="Male"
            "2"="Female"
      ;
RUN;

/* dm data set below has data only for 2 treatments and male subject is missing for
treatment 2*/

DATA dm;
      INPUT subject $ sex $ treatment $;
      CARDS;
```

```
        101 1 1
        102 2 1
        202 2 2
        203 1 1
        301 1 1
        303 2 2
;
RUN;

/*Apply format to treatment and gender*/

DATA dm1;
    SET dm;
        FORMAT treatment $trt. Sex $sex.;
        LABEL treatment="Treatment" sex="Sex";
RUN;
```

Please see below screenshot of output demographic data set. Please note that there is no observation for treatment 3. Also, male subject is not present for treatment 2.

```
    Subject    Sex        Treatment
    ---------------------------------------------
        101    Male       Treatment 1
        102    Female     Treatment 1
        203    Male       Treatment 1
        301    Male       Treatment 1
        202    Female     Treatment 2
        303    Female     Treatment 2
```

Output 1. Demographic data set

EXAMPLE 1: PROC FREQ WITH NO OPTION

Initially, to get the frequency counts by gender and treatment, we will execute PROC FREQ on the demographic data with no option.

```
/* Execute PROC FREQ to get the counts*/

PROC FREQ DATA=dm1;
      TABLES treatment*sex/ out=count (drop=percent);
RUN;
```

Please see below screenshot of output data set. Note that the following frequency counts are missing:

- For Treatment 2, frequency count for male subjects is missing.
- For Treatment 3, frequency counts for both male and female subjects are missing.

```
                            Frequency
    Treatment      Sex       Count


    ---------------------------------------------
    Treatment 1    Male         3
    Treatment 1    Female       1
    Treatment 2    Female       2
```

Output 2. Count data set when no option is used in PROC FREQ

EXAMPLE 2: PROC FREQ WITH SPARSE OPTION

In order to get the all missing frequency counts, let's try to execute PROC FREQ on the demographic data with only SPARSE option. SPARSE option reports all possible combinations of the variable values, even if a combination does not occur in the data. When you use the SPARSE and OUT= options, PROC FREQ includes empty cross tabulation table cells in the output data set. By default, PROC FREQ does not include zero-frequency table cells in the output data set.

```
/* Execute PROC FREQ with sparse option to get the counts*/

PROC FREQ DATA=dm1;
      TABLES treatment*sex/sparse out=count (drop=percent);
RUN;
```

Please see below screenshot of output data set. Note the following changes in frequency count data set.
- For Treatment 2, frequency count for male subjects is present now. Please see the highlighted row below.
- For Treatment 3, frequency counts for both male and female subjects are still missing.

```
                           Frequency
      Treatment      Sex      Count
--------------------------------------------
      Treatment 1    Male        3
      Treatment 1    Female      1
      Treatment 2    Male        0
      Treatment 2    Female      2
```

Output 3. Count data set when only SPARSE option is used in PROC FREQ

EXAMPLE 3: PROC FREQ WITH SPARSE AND PRELOADFMT (SIMULATE) OPTION

In order to get all missing frequency counts, PROC FREQ must have options similar to COMPLETETYPES and PRELOADFMT which are present in the MEANS, SUMMARY and REPORT procedures. Although, PROC FREQ has SPARSE option similar to COMPLETETYPES; however it does not have option similar to PRELOADFMT. So our best solution here is to simulate data similar to PRELOADFMT process and then execute the PROC FREQ. Please see steps below to generate PRELOADFMT option effect on demographic data:

- Create a SAS data set using PROC FORMAT with the CNTLOUT option. Then subset the data set using the treatment format. This data set will have all possible combinations of treatment.
- Merge above data set with demographic data set. The new data set will now have all possible combinations of treatment.
- Execute PROC FREQ with SPARSE option to get all possible frequency count.

Please see below SAS code for detail:

```
/*Get all possible treatments from format and rename the start variable to treatment*/

PROC FORMAT CNTLOUT=trt(KEEP=start RENAME=(start=treatment));
      SELECT $TRT;
RUN;

/*Merge all possible combination of treatment with dm*/

PROC SORT DATA = trt;
      BY treatment;
RUN;

PROC SORT DATA = dm;
      BY treatment;
RUN;
```

```
DATA dm1;
    MERGE trt dm;
    BY treatment;
        FORMAT treatment $trt. sex $sex.;
        LABEL treatment="Treatment" sex="Sex";
RUN;

PROC FREQ DATA=dm1 NOPRINT;
        TABLES treatment*sex/ sparse out=count(where=(sex~=" ") drop=percent);
RUN;
```

Please see below screenshot of output data set. Please note the following change in frequency count data set.
- For Treatment 2, frequency count for male subjects is present now. Please see the highlighted row below.
- For Treatment 3, frequency counts for both male and female subjects are present now. Please see the highlighted row below.

```
                                      Frequency
         Treatment        Sex          Count
   ----------------------------------------------
         Treatment 1      Male            3
         Treatment 1      Female          1
         Treatment 2      Male            0
         Treatment 2      Female          2
         Treatment 3      Male            0
         Treatment 3      Female          0
```

Output 4. Count data set when SPARSE and PRELOADFMT (Simulate) options are used in PROC FREQ

Overall summary, using the above method we can get the results to look as if we were using PRELOADFMT option on PROC FREQ. If multiple categories are missing then we need to merge each category from format data set separately with the original data to get the required missing frequency counts. Also, any further change in the category for e.g. an additional group added at a later stage of study will not require any change in the code.

CONCLUSION

Above solution demonstrates an innovative approach to simulate PRELOADFMT option in PROC FREQ. This solution requires minimal amount of coding and avoids any hard coding of zero count for missing category.

REFERENCES

SAS Knowledge base: http://support.sas.com/resources/

Surampalli, Naren (2012)," The power of using options COMPLETETYPES and PRELOADFMT", Proceedings of the PharmaSUG 2012 Conference,

http://www.pharmasug.org/proceedings/2012/CC/PharmaSUG-2012-CC26.pdf

DISCLAIMER

The contents of this paper are the work of the author and do not necessarily represent the opinions, recommendations, or practices of PPD.

CONTACT INFORMATION

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

Ajay Gupta, M.S
PPD
3900 Paramount Parkway
Morrisville, NC-27560
Phone: (919) -456-6461
Email: Ajay.gupta@ppdi.com
         Ajaykailasgupta@aol.com