PharmaSUG 2021 - Paper EP-190

Looking for the Missing(ness) Piece

Louise S. Hadden, Abt Associates Inc.

ABSTRACT

Reporting on missing and/or non-response data is of paramount importance when working with longitudinal surveillance, laboratory and medical record data. Reshaping the data over time to produce such statistics is a tried and true technique, but for a quick initial look at data files for problem areas, there's an easier way. This quick tip will speed up your data cleaning reconnaissance and help you find your missing(ness) piece. Additional tips on making true missingness easy to identify are included.

INTRODUCTION

There are myriad ways to determine if you have missing data (PROC FREQ, PROC MEANS, PROC SUMMARY, PROC UNIVARIATE, etc.). Most SAS® statistical procedures can report out on the number of missing values. Depending on procedure options in PROC FREQ, missing values can be included, or not, in counts of observations. Other statistical procedures simply drop records with missing values. Reporting on missing values can occur via "list" output, procedural output or ODS output objects. Most statistical procedures do not distinguish between different types of missing, but PROC FREQ and reporting procedures do. This paper and poster explore using PROC FREQ to report on missing values by variable in a data set. This presentation is suitable for all levels, industries and job roles.

PREPARING MISSING VALUES

It is common in survey output to assign codes such as 95 for other specify, 96 for other, 97 for refused, 98 for not answered, and 99 for not applicable. These values represent different types of missing. Unfortunately, to statistical procedures, they are just numbers, and are treated as such. SAS can assign up to 28 special missing value codes, ., ._, and .A through .Z. These represent extremely small numbers that are greater than 0 – different extremely small numbers. SAS can and does distinguish between these special missing values in reporting procedures and PROC FREQ. It is recommended that analysts recode the 9x codes (and the like) to special missing values – for example, 95 (other) could be .O, and a valid skip (determined by looking at a survey instrument or data dictionary for skip patterns) could be coded .V. Knowing what is truly missing data is of the utmost importance. Note information on special missing value recodes in both variable and value labels. PROC FORMAT will report on the different missing values when used with reporting procedures and PROC FREQ. When the format below is applied to a variable, .O, .M, and .V are all reported out separately.

```
proc format;
    value varx_f .O = 'Other Specify'
                 .M = 'Missing'
                 .V = 'Valid Skip'
              other = 'Non-Missing';
run;
```

PROC FREQ, ODS OUTPUT OBJECTS, AND NLEVELS

Regardless of how your data was prepared with respect to missing values, SAS has a sadly underutilized variant of PROC FREQ that allows you to produce a missingness report with ease. This procedural option is NLEVELS. If your data set has special missing numeric values, these will be reported as "levels" of missing. Character variables have a single "level" of missing. The syntax for PROC FREQ NLEVELS is as follows:

```
Ods trace;

proc freq data=int.&infi. nlevels;
  ods output nlevels=nlevels0;
```

```
  tables _all_ / noprint;
run;
ods output close;

ods trace off;

proc print data=nlevels0 (obs=5) noobs;
title 'Test nlevels output';
run;

proc contents data=nlevels0 varnum;
run;
```

The only ODS output object that PROC FREQ NLEVELS produces is NLEVELS. The NLEVELS0 data set contains 5 variables: TABLEVAR (variable name), TABLEVARLABEL (variable label), NLEVELS (number of different values, including missing), NMISSLEVELS (number of different missing values), and NNONMISSLEVELS (number of different non-missing values). Using _all_ in the tablue statement means that all variables are tabulated. Thus, you can generate a single line with missing value statistics for each variable in a single report.

PROC FREQ with the _ALL_ tables option will report on variables in the order in which they entered the PDV, so if a different order is desired, prepare your data set for reporting by reordering your variables. Additionally, ensure that all variables are labelled prior to reporting.

The listing output from PROC FREQ NLEVELS is not ideal for reporting. We use a PROC REPORT step, using the ODS OUTPUT object generated by the procedure.

PROC REPORT, ODS OUTPUT OBJECTS, AND NLEVELS

In preparation for reporting, we set the NLEVELS0 ODS output object, creating a temporary file for printing. We label the variables, and create a non-printing variable that allows us to flag any variables without any non-missing value levels.

```
data nlevels;
    set nlevels0;
    label TableVar = "Variable Name"
          TableVarLabel = "Variable Description"
          NLevels = "# of Variable values"
          NMissLevels = "# of Missing Value Levels"
          NNonMissLevels = "# of Non- Missing Value Levels";
    shadeit=(nnonmisslevels=0);
run;
```

We use PROC REPORT and ODS RTF to set up our Missingness report, highlighting a high degree of missingness using the shadeit variable created above. Note that we could modify this using cardinality ratios created from the NLEVELS, NMISSLEVELS, and NNONMISSLEVELS variables to refine our reporting. SHADEIT is set as a non-printing variable in the DEFINE statement, but must be in the columns statement in order for the conditional shading of the row to work. Note that this shading can also be applied to single cells using options in the COMPUTE statement.

```
ods rtf file="Missingingness.rtf" path=odsout style=styles.pearl;
title2 "Missingness Report for &infi - N = &nobs";
proc report nowd data=nlevels
          style(report)=[cellpadding=3pt vjust=b]
          style(header)=[just=center font_face="Helvetica" font_weight=bold
           font_size=8pt]
          style(lines)=[just=left font_face="Helvetica"] split='|';
```

```
    columns TableVar TableVarLabel NLevels NmissLevels NNonMissLevels
          shadeit;
    define shadeit / display ' ' noprint;
    define TableVar / style(COLUMN)={just=l font_face="Helvetica"
           font_size=8pt cellwidth=295 }
           style(HEADER)={just=l font_face="Helvetica" font_weight=bold
           font_size=8pt  };
    define TableVarLabel / style(COLUMN)={just=l font_face="Helvetica"
           font_size=8pt cellwidth=395 }
           style(HEADER)={just=l font_face="Helvetica" font_weight=bold
           font_size=8pt  };
    define Nlevels / style(COLUMN)={just=c font_face="Helvetica"
           foreground=navy
           font_size=8pt cellwidth=95 }
           style(HEADER)={just=c font_face="Helvetica" font_weight=bold
           font_size=8pt  };
    define NMissLevels / style(COLUMN)={just=c font_face="Helvetica"
           foreground=navy
           font_size=8pt cellwidth=95 }
           style(HEADER)={just=c font_face="Helvetica" font_weight=bold
           font_size=8pt  };
    define NNonMissLevels / style(COLUMN)={just=c font_face="Helvetica"
           foreground=navy
           font_size=8pt cellwidth=95 }
           style(HEADER)={just=c font_face="Helvetica" font_weight=bold
           font_size=8pt  };
compute shadeit;
    if (shadeit eq 1) then call
     define(_row_,"STYLE","STYLE=[BACKGROUND=PINK]");
    endcomp;
run;

ods rtf close;
```

WE FOUND THE MISSING PIECE!

We are able to produce a report from a data set with thousands of variables with a few lines of PROC FREQ and PROC REPORT code, instantly highlighting records for variables which may have a missingness problem. Using the SHADEIT variable to screen, we could produce a report with variables with only missing values to research.

Variable Name	Variable Description	# of Variable values	# of Missing Value Levels	# of Non-Missing Value Levels
IN_DATA_EXTRCT_DT	Mo 4: Date of data extraction	1	0	1
INF_IDENTIFIER1	Mo 4: Infant identifier-#1	1550	0	1550
IN_AMB_VISIT_DT	Mo 4: Date of any ambulatory care visit, including antenatal care, ED, telemedicine.	1	1	0
IN_CHLOROQ_END_DATE	Mo 4: First administration of treatment - End Date: Chloroquine Phosphate (Chloroquine)	1	0	1
IN_CHLOROQ_STRT_DATE	Mo 4: First administration of treatment - Start Date: Chloroquine Phosphate (Chloroquine)	1	0	1

Table 1. Missingness Report

CONCLUSION

PROC FREQ with the NLEVELS option can provide an excellent broad stroke report on missingness in your data sets. PROC REPORT can generate a traffic-lighted report based on the number of total levels, missing levels and non-missing levels in each individual variable in the data set. This procedural option has earned its spot in my SAS® toolkit!

REFERENCES

Boniface, Christopher J. and Wysocki, Janet L. April 2016. "You Can Bet On It, The Missing Rows are Preserved with PRELOADFMT and COMPLETETYPES". *Proceedings of the SAS Global 2016 Conference,* Las Vegas, NV: SAS Institute.

Bost, Christopher. April 2011. "To FREQ, Perchance to MEANS". *Proceedings of the SAS Global 2011 Conference,* Las Vegas, NV: SAS Institute.

Fehd, Ronald J. April 2013. "Data Review Information: N-Levels or Cardinality Ratio". *Proceedings of the SAS Global 2013 Conference,* San Francisco, CA: SAS Institute.

Jia, Justin and Lin, Amanda. April 2016. "Missing Values, They are NOT Nothing". *Proceedings of the SAS Global 2016 Conference,* Las Vegas, NV: SAS Institute.

Ramezani, Niloofar. April 2020. "Analyzing Non-normal Data: Application to Missing Data Problems". *Proceedings of the SAS Global 2020 Conference,* Virtual: SAS Institute.

Shan, Xia Ke and Bremser, Kurt. June 2020. "Five Simple Ways to Know If Variables in a Table Are All Missing". *Proceedings of the SAS Global 2020 Conference,* Virtual: SAS Institute.

Stutzman, Paul. June 2017. "Check Your Data: Tools for Automating Data Assessment". *Proceedings of the PharmaSUG 2017 Conference,* Baltimore, MD: PharmaSUG.

Zdeb, Mike. October 2016. "An Easy Route to a Missing Data Report with ODS+PROC FREQ+A Data Step". *Proceedings of the 2016 Southeast SAS Users Group Conference,* North Bethesda, MD: SESUG.

CONTACT INFORMATION

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

Louise S. Hadden
Abt Associates Inc.
Louise_hadden@abtassoc.com