

Methods to Generate Unavailable Categories in Interim Analysis

Xiaoran Yin, PPD Inc., Beijing

ABSTRACT

In most of clinical studies, we have to analyze the reasons for premature discontinuation for each treatment group. There might be 8 or more reasons be displayed on analysis table, even though some of them are unavailable then. We would like to output the unavailable count number as zero. But we can not just hard code them on our SAS programs among interim analysis period since some of these categories might be available at the final analysis. How to make our programs more powerful to handle these uncertain situations? This paper will give you methods you are looking for.

INTRODUCTION

In most of long clinical trial studies, interim analysis may be planned in protocol. It may call for once, twice, or even more iterations. But there are only a few records available in the earlier stage, such as “Reasons for Premature Discontinuation” and “AE Severity”. There are only few categories or groups available then. However, it requires the analysis report to display all of the categories even though some of them have not been available yet. The unavailable categories’ count number can be added and filled as ZERO. Once such categories become available the SAS programs have to be modified to fit the updated dataset. Since it is low efficient and erroneous, better methods shall be designed. The series of methods can avoid it and make programs more flexible and accurate. This paper gives the solutions.

REQUIRED RESULT

Reasons for Premature Discontinuation will be an example to introduce these methods. The final output required to be displayed as Table 1:

	X	Y	Z
Reasons for Premature Discontinuation:			
Subject non-compliance	xxx	xxx	xxx
Investigator/Sponsor discretion	xxx	xxx	xxx
Subject withdrew informed consent	xxx	xxx	xxx
Lost to follow-up	xxx	xxx	xxx
Other	xxx	xxx	xxx

SCENARIOS AND SOLUTIONS

1) SCENARIO 1: GENERATE A DUMMY FRAME JUST AS THE FINAL OUTPUT REQUIRED, AND THEN MERGE IT WITH THE FACTUAL ANALYSIS OUTPUT.

Assumption: the analysis dataset ADPD1 is as below:

TRTCD	SUBJID	REASONN	REASON
X	101	1	Subject non-compliance
X	102	4	Lost to follow-up
X	108	5	Other
X	110	2	Investigator/Sponsor discretion
Y	118	3	Subject withdrew informed consent
Y	124	1	Subject non-compliance
Z	136	3	Subject withdrew informed consent
Z	115	3	Subject withdrew informed consent
Z	132	1	Subject non-compliance

Solution:

```
**Step1: use "Proc freq" or "Proc sql" to calculate the available count number;
Proc freq data=adpd1;
  tables REASONNN*TRTCD /out=adpd1_v1;
run;

**Step2: generate whole categories and groups as required;
data frame;
  do REASONNN=1 to 5;
    do TRTCD="X", "Y", "Z";
      output;
    end;
  end;
run;

data adpd1_v2;
  merge frame(in=a) adpd1_v1(in=b);
  by REASONNN TRTCD;
  if a;
run;

**Step3: After that, statement "Proc transpose" can be used to display the final
result.;
Proc transpose data=adpd1_v2 out=adpd1_v3;
  by REASONNN;
  var COUNT;
  id TRTCD;
run;
```

Sometimes, we need to create a frame for multiple levels within multiple categories. Suppose that two or more key variables should be included in a frame, but these key variables separated in two existed dataset. It is difficult to create all of the categories manually, even though "do loops" were used. "Proc sql" might be a better way to create a complicated frame.

```
**Step1: the first dataset that including the key variables and unique records;
data reason;
  do REASONNN=1 to 5; output; end;
run;

**Step2: the second dataset that including the key variables and unique records;
data treatment;
  do TRTCD="X", "Y", "Z"; output; end;
run;

**Step3: merge them together and create frame;
proc sql;
  create table frame as
  select * from reason, treatment;
quit;
```

These are the general ways we always select. Is there any ways more efficient and simpler? Below is the way to omit the definition of column variables, such as the treatment groups.

2) SCENARIO 2: INITIALIZE THE UNAVAILABLE GROUPS.

Assumption: the analysis dataset is as below. Obviously, group “Z” does not available in dataset ADPD2.

TRTCD	SUBJID	REASON
X	101	Subject non-compliance
X	102	Lost to follow-up
X	108	Other
X	110	Investigator/Sponsor discretion
Y	118	Subject withdrew informed consent
Y	124	Subject non-compliance

Solution:

```
**Step1: use "proc freq" or "proc sql" to calculate the available count number;
proc freq data=adpd2;
  tables REASONNN*TRTCD /out=adpd2_v1;
run;
```

```
**Step2: transpose the available treatment groups;
proc transpose data=adpd2_v1 out=adpd2_v2;
  by REASONNN;
  var COUNT;
  id TRTCD;
run;
```

```
data frame;
  do REASONNN=1 to 5;
    output;
  end;
run;
```

**Step3: and then, we just define an array in order to initialize these 3 treatment groups;

```
data adpd2_v3;
  array trt[3] X Y Z;
  merge frame(in=a) adpd2_v2(in=b);
  by REASONNN;
  if a;
run;
```

It seemed that there wasn't any efficiency. However, in some specific situations, such as “Adverse Event”, “Medical History” and “Concomitant Medication” tables, they can benefit from this ARRAY option.

For example, in most of “Adverse Event” tables, only available preferred term categories will be required to present for each treatment groups. The only uncertain item is the treatment groups.

What shall we do now? It is similar to above scenario 2. But the process of creating frame and merging can be ignored. Only an ARRAY option can resolve all of the unavailable groups.

Another way to initialize groups is like below. You can choose the appropriate one for your study.

```
data adpd2_v3;
  X=X; Y=Y; Z=Z;
  merge frame(in=a) adpd2_v2(in=b);
  by REASONNN;
  if a;
run;
```

Is there any other way more efficient for those unavailable categories and groups? Scenario 3 will be a good choice.

3) SCENARIO 3: GENERATE FORMAT FOR ALL OF THE CATEGORIES OR GROUPS, USE COMMAND "COMPLETETYPES" AND "PRELOAD" IN "PROC SUMMARY" OR "PROC MEANS" STATEMENTS.

Assumption: the analysis dataset is same as ADPD2. Total groups that containing treatments "X", "Y" and "Z", and sub-total groups that containing treatments "X" and "Y" required to output.

Solution:

****Step1: create format for variable TRTCD and REASONN;**

```
proc format;
  value $TRTCD
    "X"="X"
    "Y"="Y"
    "Z"="Z"
  ;
  value REASONN
    1="Subject non-compliance"
    2="Investigator/Sponsor discretion"
    3="Subject withdrew informed consent"
    4="Lost to follow-up"
    5="Other"
  ;
run;

data adpd2_v1;
  format TRTCD $TRTCD. REASONN REASONN.;
  set adpd2;
  MARK=1;
run;
```

****Step2: use "proc summary" or "proc means" to calculate the available count number, and select command "COMPLETETYPES" and "PRELOADFMT".**

```
proc summary data=adpd2_v1 nway completetypes N;
  class TRTCD/preloadfmt;
  class REASONN/preloadfmt;
  var MARK;
  output out=adpd2_v2 N=count;
run;

proc sort data=adpd2_v2;
  by REASONN;
run;
```

****Step3: transpose the available treatment groups;**

```
proc transpose data=adpd2_v2 out=adpd2_v3;
  by REASONN;
  var COUNT;
  id TRTCD;
run;
```

The "COMPLETETYPES" creates all possible combinations of class variables even if the combination does not occur in the input data set.

The "PRELOADFMT" option in the CLASS statement ensures that PROC MEANS or PROC SUMMARY writes all user-defined format ranges or values for the combinations of class variables to the output, even when a frequency is zero. If the format does not been defined before PROC MEANS or PROC SUMMARY, the "PRELOADFMT" option will not be in effect.

This might be an efficient way in most of studies. It is really helpful in interim analysis. If the total column that containing all treatment groups or sub-total column required to output, the option "MULTILABEL" and "MLF" will be a much better choice. The SAS program above for step 1 and step 2 can be improved like that:

****Step1: create multiple formats for variable TRTCD;**

```
Proc format;
  value $TRTCD (multilabel)
    "X"="X"
    "Y"="Y"
    "Z"="Z"
    "X", "Y"="Sub-Total"
    "X", "Y", "Z"="Total"
  ;
  value REASONN
    1="Subject non-compliance"
    2="Investigator/Sponsor discretion"
    3="Subject withdrew informed consent"
    4="Lost to follow-up"
    5="Other"
  ;
run;

data adpd2_v1;
  format TRTCD $TRTCD. REASONN REASONN.;
  set adpd2;
  MARK=1;
run;
```

****Step2: use "Proc summary" or "Proc means" to calculate the available count number, and select command "COMPLETETYPES", "PRELOADFMT" and "MLF".**

```
Proc summary data=adpd2_v1 nway completetypes N;
  class TRTCD/preloadfmt mlf;
  class REASONN/preloadfmt ;
  var MARK;
  output out=adpd2_v2 N=count;
run;
```

"MLF" option enables PROC MEANS or PROC SUMMARY to use the primary and secondary format labels for a given range or overlapping ranges to create the subgroup combinations when a "MULTILABEL" format is assigned to a class variable.

After the statement "proc transpose", the final output will be displayed as Table 2:

	X	Y	Sub- total(X,Y)	Z	Total (X,Y,Z)
Reasons for Premature Discontinuation:					
Subject non-compliance	xxx	xxx	xxx	xxx	xxx
Investigator/Sponsor discretion	xxx	xxx	xxx	xxx	xxx
Subject withdrew informed consent	xxx	xxx	xxx	xxx	xxx
Lost to follow-up	xxx	xxx	xxx	xxx	xxx
Other	xxx	xxx	xxx	xxx	xxx

CONCLUSION

This paper provides three methods to generate unavailable categories in interim analysis. There isn't any best one but only the most appropriate one to the specific studies.

CONTACT INFORMATION

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

Name: Xiaoran Yin
Enterprise: PPD Inc.
Address: 8th.Floor, Tower B, Central Point No. 11, Dongzhimen South Ave Dongcheng District
City, State ZIP: Beijing, 100007
Work Phone: +86 10-57636250
Fax: +86 10-57636251
E-mail: Xiaoran.Yin@ppdi.com
Web: www.ppdi.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.