Paper 213-2010

Select Matched-Pair Control Sampling using PROC SURVEYSELECT

Yu Feng, IMS, Blue Bell, Pennsylvania
Paul Doucette, IMS, Blue Bell, Pennsylvania

Abstract

In marketing research, assessing the performance of a voucher/coupon program is one of the major studies. One key step for this analysis is to create a matched-pair control group, which contains the equal size observations as the categorized voucher/coupon group and the corresponding attributes of the voucher/coupon group. Stratified sampling using PROC SURVEYSELECT is a simple and efficient way to complete the task. This paper introduces and compares two approaches of SAS® programming using PROC SURVEYSELECT to select a matched-pair control group.

Introduction

One of the important studies in marketing research is to investigate the impact of the voucher program on the patients' prescription behavior. Will the voucher program have an effect on the patient adherence and persistence? In order to answer such questions, we usually select a matched-pair non-voucher control group, ideally with the equal stratum size of the voucher group. Generally in these studies, the non-voucher data set size is much larger than the voucher data set. However, that won't guarantee that each stratum sample size in the voucher data set is less than the total number of units in the stratum of the non-voucher data set. This paper presents and compares two approaches of SAS programming, both using PROC SURVEYSELECT, to select a matched-pair control sampling which corresponds to the voucher/coupon group in terms of the variables-of-interest.

Data Source and SAS Programs

In assessing the effectiveness of a voucher/coupon program, the patients are first segmented into voucher and non-voucher groups. The voucher group is the group in which we are most interested, and includes the voucher prescription data. A matched-pair control group is selected from the non-voucher group according to the stratified subgroups from the voucher group. The sample size of each stratum in the non-voucher group is equal to or less than the one in the voucher group, depending on the size differences between strata from the voucher group and the non-voucher group.

Approach 1

In this approach, we first need to set a macro variable of the record count of the voucher group data. This will be used in the later _NSIZE_ calculation for PROC SURVEYSELECT.

```
proc sql;
   select count(*) into: voucher_total
   from voucher_group
quit;
```

Next, we need to generate a data set with the distribution of the variables–of-interest in the voucher group, which also gives us the stratum size of each voucher group.

Variables description:

AGE_CAT: patient's age group;
GENDER: patient's gender;
INDEXSTRENGTH: the strength of the patient's first prescription during the study period;
INDEXSUPPLIER: the supplier ID of the patient patient's first prescription during the study period;
SPEC_GRP: specialty group;
TOT_QTY_CAT: total quantity category;

```
proc freq data= voucher_group;
tables AGE_CAT*GENDER*INDEXSTRENGTH*INDEXSUPPLIER*SPEC_GRP*TOT_QTY_CAT         /
list missing out=freq_voucher;
title "Voucher group frequency list";
run;
```

Within the distribution data for the variables-of-interest, the macro variable of the record count of the voucher group was used to calculate the sample size for the strata. Here, sample size should be greater than 0 to select samples(_NSIZE_ greater than 0).

```
DATA freq_voucher_1;
   SET freq_voucher;
   SAMPSIZE=(PERCENT*&voucher_total.)/100;
   _NSIZE_=ROUND(SAMPSIZE,1);
   IF _NSIZE_ gt 0;
RUN;
```

Then, to prepare the data sets for PROC SURVEYSELECT, we merge the non-voucher data set with the voucher frequency data set to get the strata in both of the data sets.

```
proc sort data=nonvoucher;
by AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
run;

data nonvoucher_1;
   merge nonvoucher(in=a) freq_voucher_1(in=b);
   by AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
   if a and b;
run;
```

If we don't use this step, the sampling results maybe the same from PROC SURVEYSELECT. However, if there are some strata which only exist in the non-voucher data set and are not in the voucher data set, the error notice listed below will be present in the log.

```
ERROR: The SAMPSIZE input data set does not contain a stratum to match the
current stratum of the DATA input data set.
NOTE: The above message was for the following stratum:
      Age_cat=25+ gender=Female indexstrength=40MG indexsupplier=00903
SPEC_GRP=PCP tot_qty_cat=31-60.
```

Therefore, to avoid such errors, the above merge steps need to be used.

Finally, the two input data sets are ready for PROC SURVEYSELECT. One is data set "nonvoucher_1", from which the non-voucher control sample is selected. The other is "freq_voucher_1", which includes the sample sizes for the strata (called _NSIZE_ in the data set). The simple random sampling is used (method=srs), which selects units with equal probability and without replacement. Setting the seed=123456 specifies the initial seed for random number generation. Specifying the same seed value in the PROC SURVEYSELECT reproduces the sample. The output sample data set is called "sample", and includes the sample selected from non-voucher group data set matched to the stratum in the voucher group data (freq_voucher_1).

```
proc surveyselect data=nonvoucher_1
   sampsize=freq_voucher_1
   method=srs
   seed=123456
   out=sample;
   strata AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
run;
```

A Supplement of Approach 1

If the size of every stratum in the test group (voucher group here) is greater than the corresponding size of each stratum in the control group (non-voucher group here), then it is appropriate to use this approach. Otherwise, an error will occur in the log as indicated below. The sample size in the following example is very small, but a large sample size may result in a considerable sampling error.

```
ERROR: The sample size, 3, is larger than the number of sampling units, 1.
NOTE: The above message was for the following stratum:
      Age_cat=19-24 gender=Male indexstrength=20MG indexsupplier=00037
SPEC_GRP=Other tot_qty_cat=>=91.
```

By default, PROC SURVEYSELECT does not allow us to apply a stratum sample size that is greater than the total number of units in the stratum for the without-replacement selection. As we can see from the above example, PROC SURVEYSELECT ignores such kind strata and does notselect anysample for this stratum. Fortunately, SAS 9.1(and above) allows us to change the default to SELECTALL option. Thus, we can select all stratum units even though the stratum sample size is greater than the number of units in the stratum. Following is the code:

```
proc surveyselect data=nonvoucher_1
   sampsize=freq_voucher_1
   method=srs
   seed=123456
   out=sample SELECTALL;
   strata AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
run;
```

Approach 2

To use approach 2, like approach 1, we need to generate a data set with the distribution of the variables-of-interestfor the voucher group, which also gives us each stratum size of the voucher group.

```
proc freq data=voucher_group;
tables AGE_CAT*GENDER*INDEXSTRENGTH*INDEXSUPPLIER*SPEC_GRP*TOT_QTY_CAT / list
missing out=freq_voucher;
run;
```

Instead of creating a macro variable, we run a frequency procedure for the non-voucher group data set to get the same variables-of-interest distribution as the voucher group and output to a SAS data set.

```
proc freq data=nonvoucher_group noprint;
   tables AGE_CAT*GENDER*INDEXSTRENGTH*INDEXSUPPLIER*SPEC_GRP
*TOT_QTY_CAT /list missing out=freq_nonvoucher;
run;
```

The two frequency data sets of the voucher group and non-voucher group are merged by the variables-of-interest and the resulting records are kept in both of the frequency count data sets. _NSIZE_ is the sample size for the strata and will be used later in the SAMPSIZE= of the SURVEYSELECT procedure. Choosing the minimum value between voucher count and non-voucher count guarantees sample selection even if the stratum sample size is greater than the number of units in the stratum.

```
data merge_freq;
   merge  freq_voucher(in=a drop=percent )
       freq_nonvoucher(in=b drop=percent rename=(count=nvcnt));
   by AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
   if a and b;
_nsize_=min(count, nvcnt);
run;
```

A merge step is used to get the strata in both of the data sets.

```
proc sort data=nonvoucher_group;
by AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
run;

data nonvoucher_group_1;
   merge merge_freq(in=a) nonvoucher_group(in=b);
by AGE_CAT GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
if a and b;
run;
```

The following PROC SURVEYSELECT is the same as Approach 1. Since the minimum function was used to select the minimum value between voucher count and non-voucher count in the above steps, we don't need to worry about any stratum sample being greater than the number of units in the stratum.

```
proc surveyselect data=nonvoucher_group_1
   sampsize=merge_freq
   method=srs
   seed=123456
   out=temp;
   strata AGE_CAT  GENDER INDEXSTRENGTH INDEXSUPPLIER SPEC_GRP TOT_QTY_CAT;
run;
```

Conclusion

PROC SURVEYSELECT is a simple and efficient method to select matched-pair control stratified sampling. Both of the approaches presented here can help prepare input data sets for PROC SURVEYSELECT.

References

SAS Online Documentation, version 9.1, SAS/STAT User's Guide, SAS Institute.

Contact Information

Yu Feng
Statistician
IMS
960A Harvest Drive
Blue Bell, PA 19422
yfeng@us.imshealth.com

Paul Doucette
Senior Manager
IMS
960A Harvest Drive
Blue Bell, PA 19422
pdoucette@us.imshealth.com