SAS® GLOBAL FORUM 2020 #SASGF

Paper 4671-2020

Synchronized Multivariate Resampling to Designated Distribution and Population Level with PROC SURVEYSELECT

Zhiyong Chen, Zem Data Science LLC, Previous Pharmerit International Inc

*The intellectual property belongs to Pharmerit International Inc

ABSTRACT

PROC SURVEYSELECT is a powerful SAS® procedure for random resampling. With SAMPSIZE and STRATA option, the population level can be altered in the resampled data for designated variables. To further extend the function of PROC SURVEYSELECT, we developed an innovative approach which can perform synchronized multivariate resampling with PROC SURVEYSELECT. The approach first prepares a cross-bin flag through crossing all involved variables, then calculates the expected percent for each level of the created cross-bin flag by crossing the designated percent of each level in each involved variable. Based on the derived percent, the sample number for each cross-bin level is calculated, and finally applied in PROC SURVEYSELECT for resampling.

INTRODUCTION

Although a data collected by random sampling can deliver certain level precise inferences of the whole population, it could introduce bias to the inferences. In addition, the sampled data can only be applied in deducting a single estimate, with little information on the variability or uncertainty in the estimate. Thus, resampling of the original sample data is necessary so the inferences at different level population parameter can be concluded (Brownlee 2018). In Matching-Adjusted Indirect Comparison (MAIC) Analysis, resampling, especially multivariate resampling, is critical when the historical patient level data is not available (Malangone and Sherman 2011).

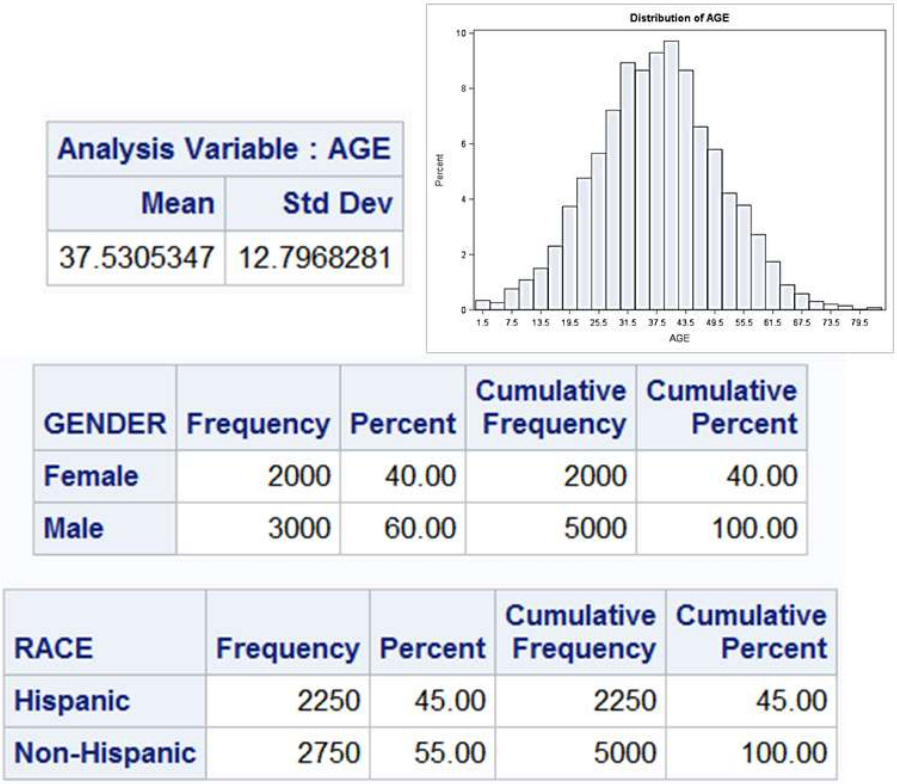
In SAS, resampling can be performed with BOOTSTRAP code (Cassell 2010) or PROC SURVEYSELECT step (Bordenave 2015). In this manuscript, an innovative approach to perform synchronized multivariate resampling with PROC SURVEYSELECT is introduced.

SIMULATED DEMO DATA

The follow DATA step code creates a simulated demo data file named ORIGINAL_DATA which has 5000 observations, with 40% Female and 60% Male, 45% Hispanic and 55% Non-Hispanic, and average age around 37.5 year old.

```
DATA ORIGINAL_DATA (DROP = I);
   LENGTH GENDER $6 RACE $15;
   CALL STREAMINIT(3);
   DO I = 1 TO 3000;
         GENDER = 'Male';
         IF I LE 1350 THEN RACE = 'Hispanic';
         ELSE RACE = 'Non-Hispanic';
         AGE = RAND("NORMAL", 37.5, 13);
         IF AGE < 0 THEN AGE = 1;
         OUTPUT;
   END;
   DO I = 1 TO 2000;
         GENDER = 'Female';
         IF I LE 900 THEN RACE = 'Hispanic';
         ELSE RACE = 'Non-Hispanic';
         AGE = RAND("NORMAL", 37.5, 13);
         IF AGE < 0 THEN AGE = 1;
         OUTPUT;
   END;
RUN;
```

PROC MEANS, PROC UNIVARIATE, and PROC FREQ confirmed the distribution and population of the variables (age, gender and race) in the simulated demo data (Figure 1).



Analysis Variable : AGE	
Mean	Std Dev
37.5305347	12.7968281

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2000	40.00	2000	40.00
Male	3000	60.00	5000	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hispanic	2250	45.00	2250	45.00
Non-Hispanic	2750	55.00	5000	100.00

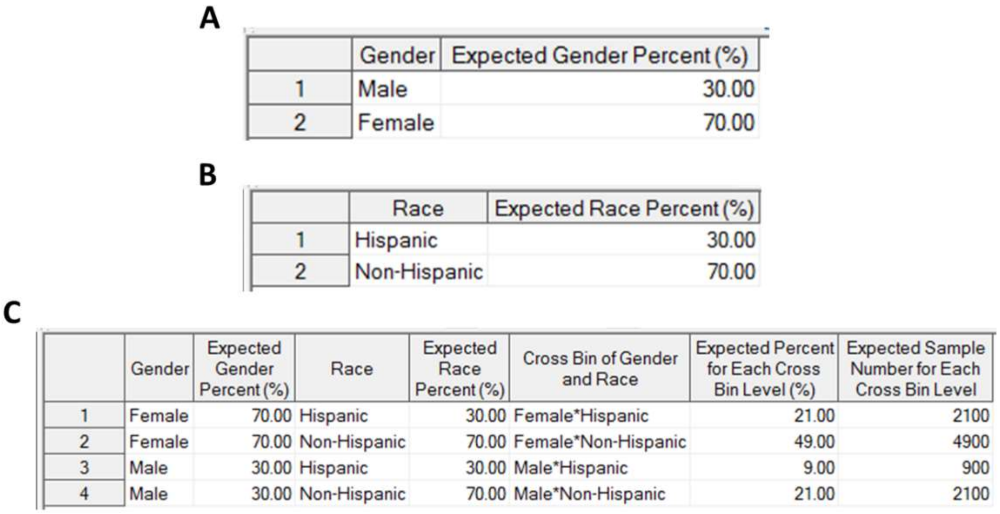
Figure 1 Distribution of Age and Population Levels pf Gender and Race in the simulated Data

RESAMPLING BASED ON DESIGNATED CATEGORICAL VARIABLES' POPULATION

In a resampling case which expects 30% Male, 70% Female (Figure 2A), and 30% Hispanic, 70% Non-Hispanic (Figure 2B), after gender and race are crossed with the following PROC SQL step for cartesian join,

```
PROC SQL NOPRINT;
   CREATE TABLE EXPECT_GENDER_RACE AS
   SELECT A.*, B.*, CATX('*', A.GENDER, B.RACE) AS CROSS_BIN LABEL =
         'Cross Bin of Gender and Race',
         EXPECT_GENDER*B.EXPECT_RACE/100 AS EXPECT_PERCENT FORMAT = 8.2
         LABEL = 'Expected Percent for Each Cross Bin Level',
         ROUND(10000*CALCULATED EXPECT_PERCENT/100) AS EXPECT_SAMPLENUM
         LABEL = 'Expected Sample Number for Each Cross Bin Level'
   FROM EXPECT_GENDER AS A, EXPECT_RACE AS B
   ORDER BY CALCULATED CROSS_BIN;
QUIT;
```

the percent numbers are 9% (30% × 30%) for Male Hispanic, 21% (70% × 30%) for Female Hispanic, 21% (30% × 70%) for Male Non-Hispanic, and 49% (70% × 70%) for Female Non-Hispanic. If 10000 observations are expected, the observation numbers for each category are 900 (10000 × 9%), 2100 (10000 × 21%), 2100 (10000 × 21%), and 4900 (10000 × 49%) respectively (Figure 2C).



	Gender	Expected Gender Percent (%)
1	Male	30.00
2	Female	70.00

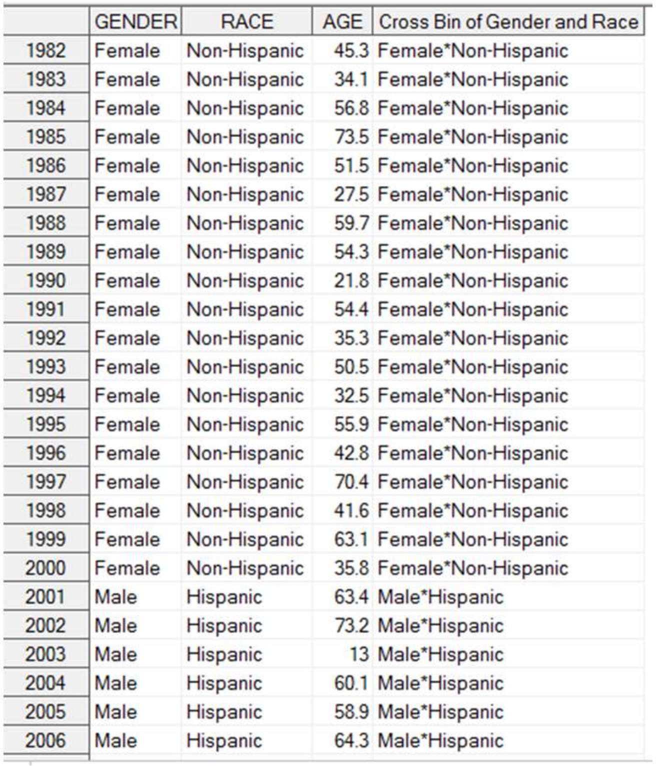
	Race	Expected Race Percent (%)
1	Hispanic	30.00
2	Non-Hispanic	70.00

	Gender	Expected Gender Percent (%)	Race	Expected Race Percent (%)	Cross Bin of Gender and Race	Expected Percent for Each Cross Bin Level (%)	Expected Sample Number for Each Cross Bin Level
1	Female	70.00	Hispanic	30.00	Female*Hispanic	21.00	2100
2	Female	70.00	Non-Hispanic	70.00	Female*Non-Hispanic	49.00	4900
3	Male	30.00	Hispanic	30.00	Male*Hispanic	9.00	900
4	Male	30.00	Non-Hispanic	70.00	Male*Non-Hispanic	21.00	2100

Figure 2 Expected Population Levels of Gender and Race after Resampling. (A) Expected gender percent; (B) Expected race percent; (C) Expected percent and sample number for each cross bin after gender and race are crossed

Before PROC SURVEYSELECT step, a cross-bin flag between GENDER and RACE for each observation in the simulated demo file needs to be prepared. The following PROC SQL creates the flag and orders the data based on the created flag (Figure 3).

```
PROC SQL;
   CREATE TABLE WITH_CROSS_BIN AS
   SELECT *, CATX('*', GENDER, RACE) AS CROSS_BIN LENGTH = 20 LABEL =
         'Cross Bin of Gender and Race'
   FROM ORIGINAL_DATA
   ORDER BY CALCULATED CROSS_BIN;
QUIT;
```



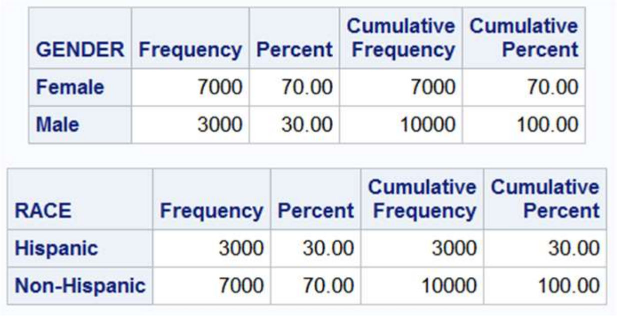
	GENDER	RACE	AGE	Cross Bin of Gender and Race
1982	Female	Non-Hispanic	45.3	Female*Non-Hispanic
1983	Female	Non-Hispanic	34.1	Female*Non-Hispanic
1984	Female	Non-Hispanic	56.8	Female*Non-Hispanic
1985	Female	Non-Hispanic	73.5	Female*Non-Hispanic
1986	Female	Non-Hispanic	51.5	Female*Non-Hispanic
1987	Female	Non-Hispanic	27.5	Female*Non-Hispanic
1988	Female	Non-Hispanic	59.7	Female*Non-Hispanic
1989	Female	Non-Hispanic	54.3	Female*Non-Hispanic
1990	Female	Non-Hispanic	21.8	Female*Non-Hispanic
1991	Female	Non-Hispanic	54.4	Female*Non-Hispanic
1992	Female	Non-Hispanic	35.3	Female*Non-Hispanic
1993	Female	Non-Hispanic	50.5	Female*Non-Hispanic
1994	Female	Non-Hispanic	32.5	Female*Non-Hispanic
1995	Female	Non-Hispanic	55.9	Female*Non-Hispanic
1996	Female	Non-Hispanic	42.8	Female*Non-Hispanic
1997	Female	Non-Hispanic	70.4	Female*Non-Hispanic
1998	Female	Non-Hispanic	41.6	Female*Non-Hispanic
1999	Female	Non-Hispanic	63.1	Female*Non-Hispanic
2000	Female	Non-Hispanic	35.8	Female*Non-Hispanic
2001	Male	Hispanic	63.4	Male*Hispanic
2002	Male	Hispanic	73.2	Male*Hispanic
2003	Male	Hispanic	13	Male*Hispanic
2004	Male	Hispanic	60.1	Male*Hispanic
2005	Male	Hispanic	58.9	Male*Hispanic
2006	Male	Hispanic	64.3	Male*Hispanic

Figure 3 Samples of the Cross-bin Flag

An Unrestricted Random (with equal probability and replacement) resampled data can now be prepared by the following code.

```
PROC SURVEYSELECT DATA = WITH_CROSS_BIN SEED = 1234 METHOD = URS OUTHITS
                  OUT = REPLACEMENT_GENDER_RACE (DROP = CROSS_BIN)
   SAMPSIZE = (2100, 4900, 900, 2100);
   STRATA CROSS_BIN;
RUN;
```

PROC FREQ test of the resampled data (Figure 4) displayed that the percent for both GERDER and RACE are exactly the same as the expected ones.



GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	7000	70.00	7000	70.00
Male	3000	30.00	10000	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hispanic	3000	30.00	3000	30.00
Non-Hispanic	7000	70.00	10000	100.00

Figure 4 Population Level of Gender and Race after Unrestricted Random Resampling

For simple random (with equal probability and without replacement) resampling, two extra steps are needed in determining the sample size.

First, the percent of the cross-bin flag in ORIGINAL_DATA is compared with the ones listed in Figure 2 using the following PROC SQL code, and the original sample number for the bin with the maximal increasing percent is chosen for the following calculation, because in theory even all observations at that level are chosen into the resampled data, other bin level should still have some extra samples. In this case, 1100 for Female*Non-Hispanic bin is chosen (Figure 5A).

```
PROC SQL;
   CREATE TABLE COMPARE AS
   SELECT A.*, B.EXPECT_PERCENT,
         EXPECT_PERCENT - A.ORIGINAL_PERCENT AS PERCENT_CHANGE
         LABEL = 'Increase Percent Number after Resampling'
   FROM ORIGINAL_SUMMARY AS A, EXPECT_GENDER_RACE AS B
   WHERE A.CROSS_BIN = B.CROSS_BIN;
QUIT;
```

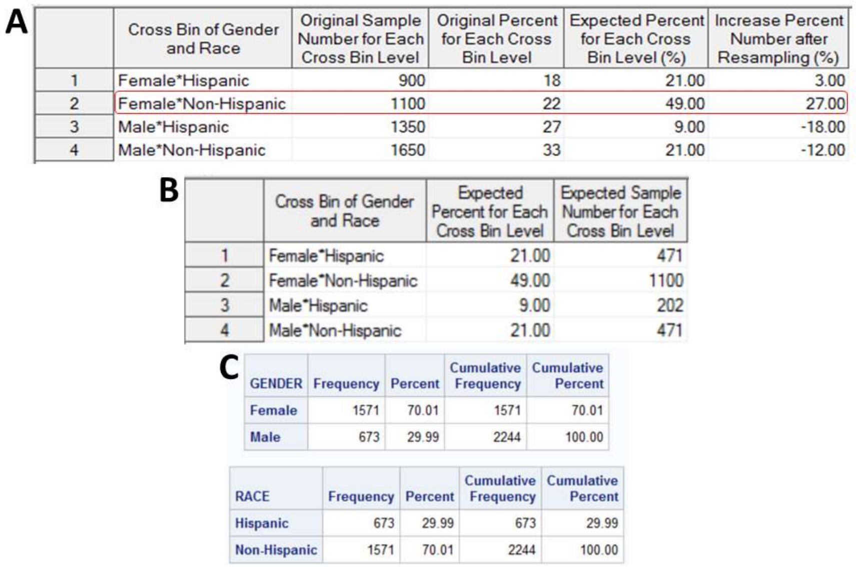
Sample number of other cross-bin level can then be calculated using equation of chosen sample number (1100)/chosen percent(49)*the expected percent at each cross-bin level (Figure 5B),

```
 PROC SQL;
   CREATE TABLE EXPECT_GENDER_RACE AS
   SELECT CROSS_BIN, EXPECT_PERCENT, ROUND(1100/49*EXPECT_PERCENT) AS
         EXPECT_SAMPLENUM LABEL = 'Expected Sample Number for Each Cross
         Bin Level'
   FROM COMPARE;
 QUIT;
```

and the final Simple Random resampled data can be prepared by the following code.

```
PROC SURVEYSELECT    DATA = WITH_CROSS_BIN SEED = 1234 METHOD = SRS
               OUT = REPLACEMENT_GENDER_RACE (DROP = CROSS_BIN)
   SAMPSIZE = (471, 1100, 202, 471);
   STRATA CROSS_BIN;
RUN;
```

PROC FREQ test of the resampled data (Figure 5C) displayed that the percent for both GERDER and RACE are close to the expected ones.



A

	Cross Bin of Gender and Race	Original Sample Number for Each Cross Bin Level	Original Percent for Each Cross Bin Level	Expected Percent for Each Cross Bin Level (%)	Increase Percent Number after Resampling (%)
1	Female*Hispanic	900	18	21.00	3.00
2	Female*Non-Hispanic	1100	22	49.00	27.00
3	Male*Hispanic	1350	27	9.00	-18.00
4	Male*Non-Hispanic	1650	33	21.00	-12.00

B

	Cross Bin of Gender and Race	Expected Percent for Each Cross Bin Level	Expected Sample Number for Each Cross Bin Level
1	Female*Hispanic	21.00	471
2	Female*Non-Hispanic	49.00	1100
3	Male*Hispanic	9.00	202
4	Male*Non-Hispanic	21.00	471

C

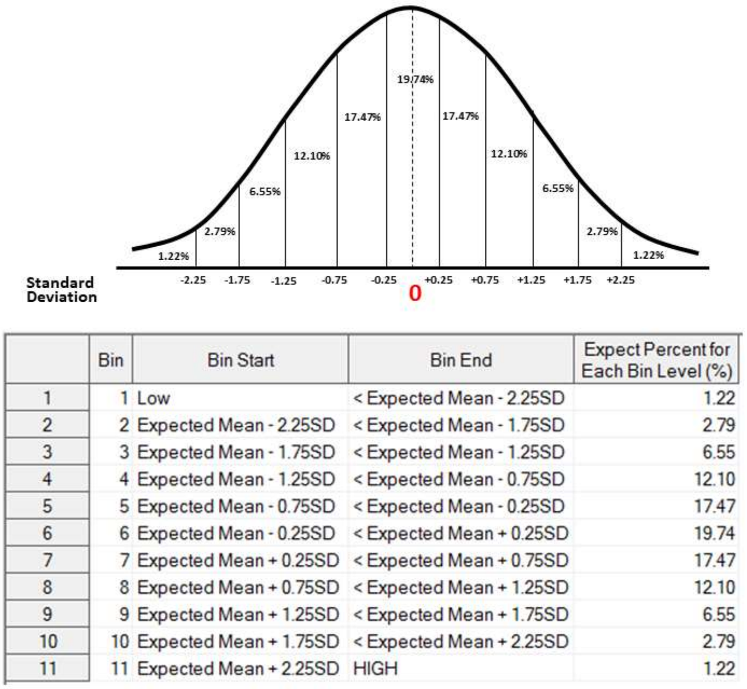
GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	1571	70.01	1571	70.01
Male	673	29.99	2244	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hispanic	673	29.99	673	29.99
Non-Hispanic	1571	70.01	2244	100.00

Figure 5 Sample Number Calculation before and Population Level after Simple Random Resampling. (A) Comparation between original and expected percent of each cross bin; (B) Sample number calculated based on the expected sample size and expected cross bin percent; (C) PROC FREQ result of gender after resampling; (D) PROC FREQ result of race after resampling

ADD NUMERICAL VARIABLE(S) TO RESAMPLING

Numerical variables first need to be formatted into 11 bin levels, whose percent is determined based on the Standard Normal Distribution Table (https://www.mathsisfun.com). The formats are prepared based on Standard Normal Distribution Table as well, using the expected mean value, and the SD derived from the original data (Figure 6).



	Bin	Bin Start	Bin End	Expect Percent for Each Bin Level (%)
1	1	Low	< Expected Mean - 2.25SD	1.22
2	2	Expected Mean - 2.25SD	< Expected Mean - 1.75SD	2.79
3	3	Expected Mean - 1.75SD	< Expected Mean - 1.25SD	6.55
4	4	Expected Mean - 1.25SD	< Expected Mean - 0.75SD	12.10
5	5	Expected Mean - 0.75SD	< Expected Mean - 0.25SD	17.47
6	6	Expected Mean - 0.25SD	< Expected Mean + 0.25SD	19.74
7	7	Expected Mean + 0.25SD	< Expected Mean + 0.75SD	17.47
8	8	Expected Mean + 0.75SD	< Expected Mean + 1.25SD	12.10
9	9	Expected Mean + 1.25SD	< Expected Mean + 1.75SD	6.55
10	10	Expected Mean + 1.75SD	< Expected Mean + 2.25SD	2.79
11	11	Expected Mean + 2.25SD	HIGH	1.22

Figure 6 Standard Normal Distribution Table

For AGE variable in the simulated demo file, the following step is applied for a format, with the value listed in the format calculated based on the expected age after resampling (presumably 47.5) and SD (12.80) listed in Figure 1.

```
PROC FORMAT;
   VALUE AGEFORMAT
         LOW   - <18.71    = AGE01
         18.71 - <25.11    = AGE02
         25.11 - <31.50    = AGE03
         31.50 - <37.90    = AGE04
         37.90 - <44.30    = AGE05
         44.30 - <50.70    = AGE06
         50.70 - <57.10    = AGE07
         57.10 - <63.50    = AGE08
         63.50 - <69.89    = AGE09
         69.89 - <76.29    = AGE10
         76.29 - HIGH      = AGE11;
RUN;
```

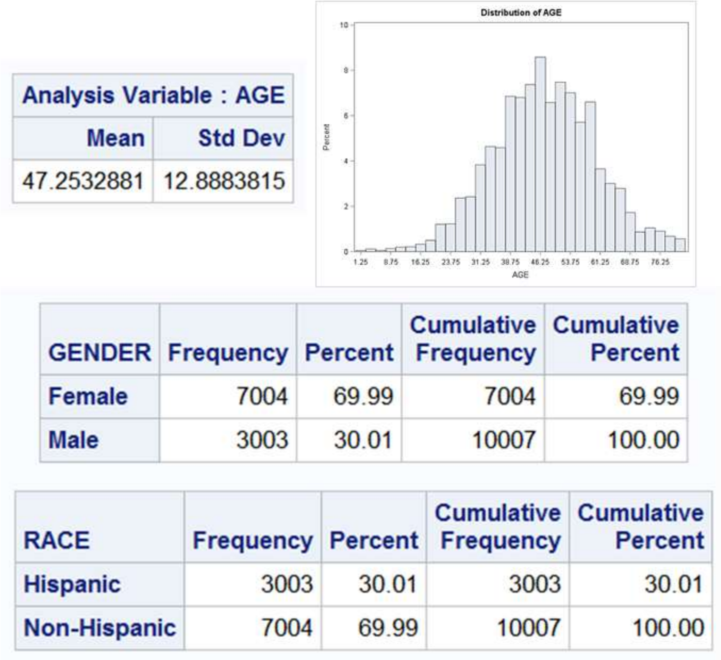
With the same approach for GENDER and RACE, both replacement or non-replacement resampling can be performed on GENDER, RACE, and AGE with the following code. Because there are 2 (for GENDER) × 2 (for RACE) × 11 (for AGE) = 44 cross-bin levels in total, a macro value which holds all sample numbers for all cross-bin level is used in PROC SURVEYSELECT.

```
*Expected percent of crossed gender, race and age after resampling;
PROC SQL NOPRINT;
   CREATE TABLE EXPECT_GENDER_RACE_AGE AS
   SELECT A.*, B.*, CATS('AGE', PUT(C.GRP, Z2.)) AS AGE, C.PERCENT AS
         EXPECT_AGE LABEL = 'Expect Percent for Age Bin Level (%)',
         CATX('*', A.GENDER, B.RACE, CALCULATED AGE) AS CROSS_BIN
         LENGTH = 500 LABEL = 'Cross Bin of Gender, Race and Age',
         EXPECT_GENDER*B.EXPECT_RACE*C.PERCENT/100/100 AS EXPECT_PERCENT
         FORMAT=8.2 LABEL='Expected Percent for Each Cross Bin Level (%)',
         ROUND(10000*CALCULATED EXPECT_PERCENT/100) AS EXPECT_SAMPLENUM
         LABEL = 'Expected Sample Number for Each Cross Bin Level'
   FROM EXPECT_GENDER AS A, EXPECT_RACE AS B, NORMAL_PERCENT AS C
   ORDER BY CALCULATED CROSS_BIN;

   SELECT EXPECT_SAMPLENUM INTO: EXPECT_SAMPLENUM SEPARATED BY ","
   FROM EXPECT_GENDER_RACE_AGE;

   CREATE TABLE WITH_CROSS_BIN AS
   SELECT *, CATX('*', GENDER, RACE, PUT(AGE, AGEFORMAT.)) AS CROSS_BIN
         LENGTH = 500 LABEL = 'Cross Bin of Gender, Race and Age'
   FROM ORIGINAL_DATA
   ORDER BY CALCULATED CROSS_BIN;
QUIT;

*Replacement resampling;
PROC SURVEYSELECT DATA = WITH_CROSS_BIN SEED = 1234 METHOD = URS OUTHITS
               OUT = REPLACEMENT_GENDER_RACE_AGE
   SAMPSIZE = (&EXPECT_SAMPLENUM);
   STRATA CROSS_BIN;
RUN;

*Non-replacement resampling;
PROC FREQ DATA = WITH_CROSS_BIN;
```

```
   TABLE CROSS_BIN / OUT = ORIGINAL_SUMMARY (RENAME = (COUNT =
   ORIGINAL_SAMPLENUM PERCENT = ORIGINAL_PERCENT));
RUN;

PROC SQL NOPRINT;
   CREATE TABLE COMPARE AS
   SELECT A.*, B.EXPECT_PERCENT, B.EXPECT_PERCENT - A.ORIGINAL_PERCENT AS
         PERCENT_CHANGE FORMAT = 8.2 LABEL = 'Increase Percent Number
         after Resampling (%)'
   FROM ORIGINAL_SUMMARY AS A, EXPECT_GENDER_RACE_AGE AS B
   WHERE A.CROSS_BIN = B.CROSS_BIN;

   CREATE TABLE EXPECT_GENDER_RACE_AGE AS
   SELECT CROSS_BIN, EXPECT_PERCENT, ORIGINAL_SAMPLENUM,
         CASE  WHEN ROUND(95/6.66*EXPECT_PERCENT) > ORIGINAL_SAMPLENUM
               THEN ORIGINAL_SAMPLENUM
               ELSE ROUND(95/6.66*EXPECT_PERCENT) END AS EXPECT_SAMPLENUM
         LABEL = 'Expected Sample Number for Each Cross Bin Level'
   FROM COMPARE
   ORDER BY CROSS_BIN;

   SELECT EXPECT_SAMPLENUM INTO: EXPECT_SAMPLENUM SEPARATED BY ","
   FROM EXPECT_GENDER_RACE_AGE;
QUIT;

PROC SURVEYSELECT    DATA = WITH_CROSS_BIN SEED = 1234 METHOD = SRS
               OUT = NON_REPLACEMENT_GENDER_RACE_AGE
   SAMPSIZE = (&EXPECT_SAMPLENUM);
   STRATA CROSS_BIN;
RUN;
```

In non-replacement resampling, it is possible that some level(s) might have less observations than the calculated ones. In the case, the sample number in the original data is chosen instead of the calculated ones (the CASE statement when creating EXPECT_GENDER_RACE_AGE). Age distribution and population level of gender and race after resampling are displayed in Figure 7 (for replacement resampling) and Figure 8 (for non-replacement resampling).

In case there are missing cross-bin level in the original data, the expected percent for a missing bin level can be either added to the adjacent ones, combining together (i.e. combining levels less than 1% into a single level), or the expected sample number can be recalibrated by dividing the total of all non-missing cross-bin level percent, which should be less than 100% after the percent of missing is excluded.

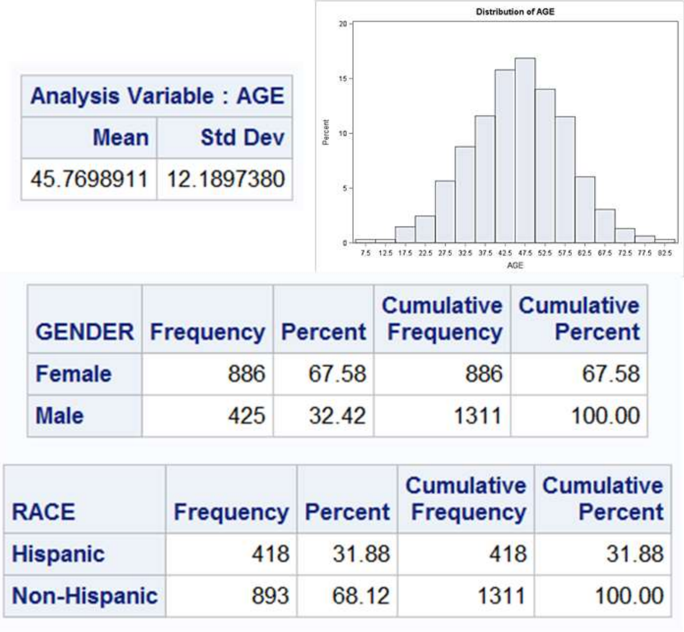


Analysis Variable : AGE	
Mean	Std Dev
47.2532881	12.8883815

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	7004	69.99	7004	69.99
Male	3003	30.01	10007	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hispanic	3003	30.01	3003	30.01
Non-Hispanic	7004	69.99	10007	100.00

Figure 7 Distribution of Age and Population Level of Gender/Race after Unrestricted Random Resampling



Analysis Variable : AGE	
Mean	Std Dev
45.7698911	12.1897380

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	886	67.58	886	67.58
Male	425	32.42	1311	100.00

RACE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hispanic	418	31.88	418	31.88
Non-Hispanic	893	68.12	1311	100.00

Figure 8 Distribution of Age and Population Level of Gender/Race after Simple Resampling

CONCLUSION

Applying cross-bin approach with PROC SURVEYSELECT provides a reliable way to perform synchronized multivariate resampling, which can deliver expected distribution/population level simulated data through both unrestricted random and simple random resampling.

REFERENCES

Brownlee J. 2018. “A Gentle Introduction to Statistical Sampling and Resampling.” Accessed June 13, 2018. https://machinelearningmastery.com/statistical-sampling-and-resampling/.

Malangone E. and Sherman S. 2011. “Matching-Adjusted Indirect Comparison Analysis Using Common SAS® 9.2 PROCEDURES.” Proceedings of the SAS Global 2011. Las Vegas, NV: SAS Institute Inc.

Cassel, D. 2010. “BootstrapMania!: Re-Sampling the SAS® Way.” Proceedings of the SAS Global 2010. Seattle, WA: SAS Institute Inc.

Bordenave, R. 2015. “Using PROC SURVEYSELECT: Random Sampling.” Proceedings of the Southeast SAS Users Group Conference 2015. Savannah, GA: SouthEast SAS® Users Group.

CONTACT INFORMATION

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

Zhiyong Chen
Zem Data Science, LLC
571-215-0014
Zhiyongchen03@yahoo.com
www.zemdata.com