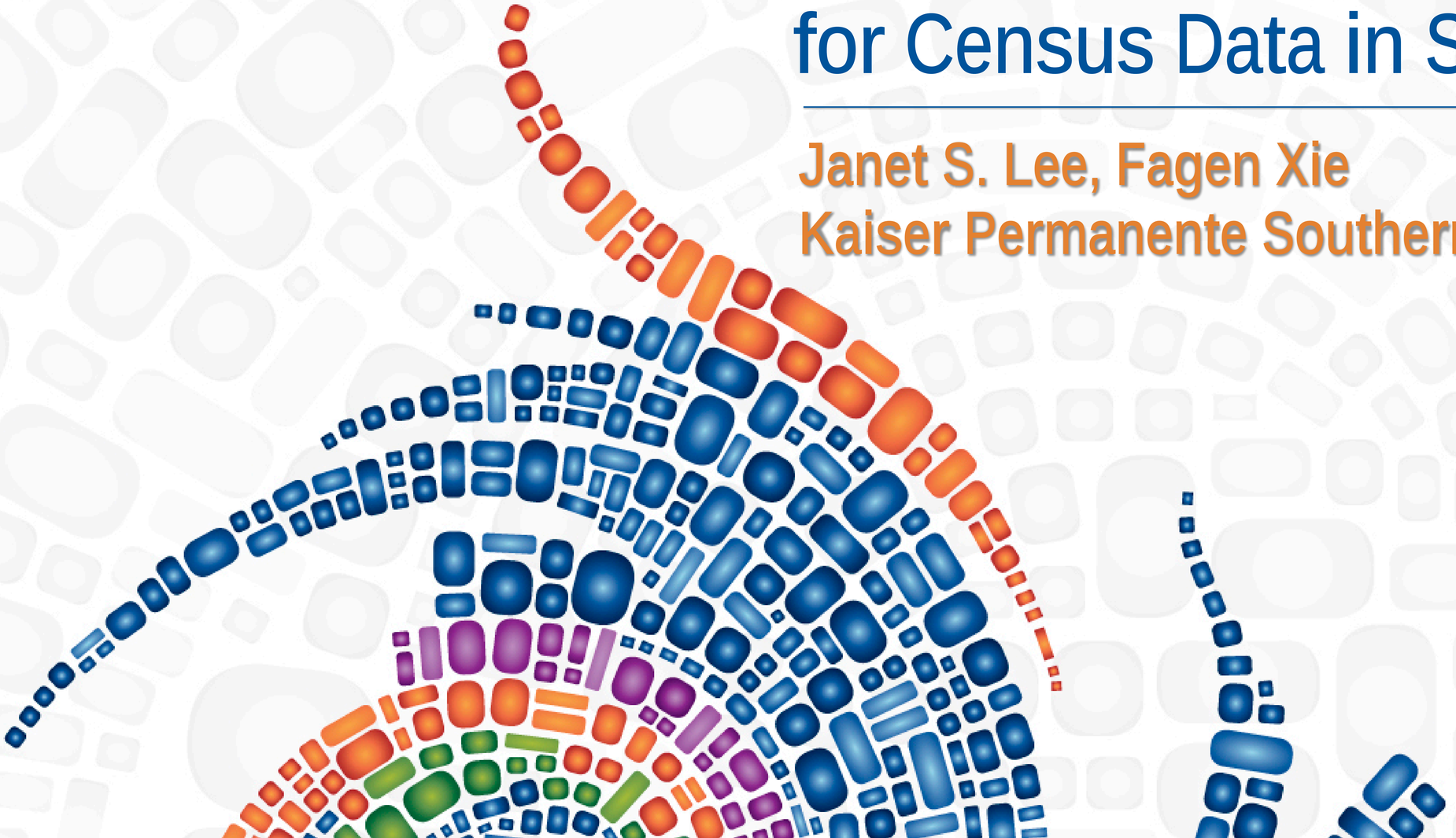
SAS® GLOBAL FORUM 2015

The Journey Is Yours

Automation of Statistics Summary for Census Data in SAS®

Janet S. Lee, Fagen Xie
Kaiser Permanente Southern California



SAS GLOBAL FORUM

Automation of Statistics Summary for Census Data in SAS®

Janet S. Lee, Fagen Xie
Kaiser Permanente Southern California

Abstract

Census data, such as education and income, has been extensively used for various purposes. The data is usually collected in percentages of census unit levels based on the population sample. Such presentation of the data makes it hard to interpret and compare. A more convenient way of presenting the data is to use the geocoded percentage to produce counts for a pseudo-population. We developed a very flexible SAS® macro to automatically generate the descriptive summary tables for the census data as well as conduct statistical tests to compare the different levels of the variable by groups. The SAS® macro is not only useful for census data but can be used to generate summary tables for any data with percentages in multiple categories.

Objectives

- Automate summarization of census data into a table with appropriate comparisons by using a SAS® macro

Methods

- Below is an example of a dataset with 3 levels for education that describe the percent of residents within the person's census block that fall into each education category. Note that the total of all 3 categories sum to 100.

ID	education1	education2	education3
A	16.4	43.4	40.2
B	20.8	39.8	39.4
C	75.6	19	5.4
⋮			

- Specify input dataset, list of geocoded variables, the label for the geocoded data, percent type (optional), percent decimal, and whether missing values should be considered a distinct category

```
%Geocode_table(Dataset=geocode_data,
Varlist=EDUCATION1 EDUCATION2
EDUCATION3,
Varlabel=Education (Pop 25+),
pcttype=col,
decpct=0,
incmiss=Y);
```

Results

- The macro outputs the results into a formatted table . The macro allows for a variety of ways to style the output to fit .

	Total (N=504)
Education (Pop 25+)[1]	
High school graduate or less	170 (34%)
Some college, Associates Degree	150 (30%)
Bachelor's degree or higher	184 (37%)

(report generated on 24AUG2014)

[1] The counts are not for the actual cohort but are estimated based on average percent.

Conclusions

- The macro %Geocode_table allows users to take geocoded data and produce a table that can be directly inserted into a report or manuscript.

References

- Census 2010. http://www.census.gov/2010census/
- Novotny PJ, Tan, AD, Foster NR, Sloan JA. (2012) SAS Tools for Cost Effective and High Quality Clinical Trial Reporting, 345-2012, Proceeding of SAS Global Forum. Available for download at: http://support.sas.com/resources/papers/proceedings12/345-2012.pdf.

Janet S. Lee, Fagen Xie
Kaiser Permanente Southern California

ADDITIONAL EXAMPLES

•What if we are interested in looking at the data by group?

ID	Group	education1	education2	education3
A	1	16.4	43.4	40.2
B	1	20.8	39.8	39.4
C	2	75.6	19	5.4
⋮				

•Include the group variable of interest into the macro

```
%Geocode_table(Dataset=geocode_data,
Varlist=EDUCATION1 EDUCATION2 EDUCATION3,
Varlabel=Education (Pop 25+),
Group=Group,
pcttype=col,
decpct=0,
incmiss=Y);
```

Macro Step 1: Run PROC MEANS for each geocoded variable to obtain mean percent for each category

Macro Step 2: Create pseudo counts by multiplying the mean percent with the total counts for each group in a DATA step.

	Group 1 (N=340)	Group 2 (N=164)	Total (N=504)
Education (Pop 25+)[1]			
High school graduate or less	108 (32%)	62 (38%)	170 (34%)
Some college, Associates Degree	107 (31%)	43 (26%)	150 (30%)
Bachelor's degree and higher	125 (37%)	59 (36%)	184 (37%)

(report generated on 24AUG2014)

[1] The counts are not for the actual cohort but are estimated based on average percent.

•The macro can also run a test to evaluate whether the counts differ by group and will include a p-value in the table.

Macro Step 3 (Optional): Use PROC FREQ to run the chi-square test on the estimated counts from Step 2 to test for comparison by group

```
%Geocode_table(Dataset=geocode_data,
Varlist=EDUCATION1 EDUCATION2
EDUCATION3,
Varlabel=Education (Pop 25+),
Group=Group,
Grouptest=Y,
pcttype=col,
decpct=0,
incmiss=Y);
```

	Group 1 (N=340)	Group 2 (N=164)	Total (N=504)	p value
Education (Pop 25+)[1]				0.326
High school graduate or less	108 (32%)	62 (38%)	170 (34%)	
Some college, Associates Degree	107 (31%)	43 (26%)	150 (30%)	
Bachelor's degree and higher	125 (37%)	59 (36%)	184 (37%)	

(report generated on 24AUG2014)

[1] The counts are not for the actual cohort but are estimated based on average percent.

Janet S. Lee, Fagen Xie
Kaiser Permanente Southern California

LIMITATIONS

1. Sample size

•Approximation of cohort counts using summary estimates is a reasonable to do when the total counts are large enough. Each person is representative of the demographic distribution of its census block so having a small sample size may result in estimated counts that aren't very informative.

- Example where N=3

ID	education1	education2	education3
A	16.4	43.4	40.2
B	20.8	39.8	39.4
C	75.6	19	5.4

	Total (N=3)
Education (Pop 25+)[1]	
High school graduate or less	1 (33%)
Some college, Associates Degree	1 (33%)
Bachelor's degree or higher	1 (33%)

(report generated on 24AUG2014)

[1] The counts are not for the actual cohort but are estimated

•Another issue that may be tied to sample size is having too many categories. There are 14 distinct education levels captured by the U.S. Census and not all of those categories may be relevant for analysis so it may be helpful to group them into more meaningful categories (as was done in previous examples).

- For example, the following categories: Nursery School, Kindergarten, Grade 1 through 11, 12th grade no diploma could be combined into one category "Less than high school". If a study was being conducted in an area where these distinct categories may have few, if any, people, having such a level of detail would be unnecessary.

2. Alternative data

- Use the actual variable of interest if it has been captured for most individuals. The actual variable will provide more information than summary level percents from geocoded data. If the variable is to be used in a model, it will be easier to interpret the effects of the actual variable rather than the geocoded percents.

- Some analyses may be more concerned with neighborhood SES. In those cases, creating an indicator variable may be more useful.
  - For example, an indicator for neighborhoods with household income<FPL may be more useful than using different levels of household income.

SAS® GLOBAL FORUM

April 26-29
Dallas, TX