

Paper 120

Many Tables, Short Turnaround: Displaying Survey Results Quickly and Efficiently

Rebecca V. Fink, Abt Associates; David Izrael, Abt Associates; Sarah W. Ball, Abt Associates; Sara M. A. Donahue, Abt Associates

ABSTRACT

When constructing multiple tables with a tight deadline, it is critical to have a tool to quickly and efficiently create tables using a standard set of parameters. We demonstrate a set of SAS® macros based on PROC SURVEYFREQ which can be used to summarize survey data in tables that contain unweighted and weighted counts and percentages, as well as a weighted treatment ratio, with respective confidence intervals. These macros allow us to display both single-select survey questions (i.e., survey questions with only one response allowed, such as gender) and multiple choice survey questions that allow the respondent to choose more than one response (for example, insurance status) within the same table. Further, we can use these macros to sort the output by the treatment ratio or by the column percent distribution in either ascending or descending order. Finally, these macros have the ability to output the results on a defined subset of the sample.

INTRODUCTION

This paper builds upon the macros described in a previous paper [1].

Our assumptions are as follows:

- 1) The survey data includes a calculated survey weight and a variable to identify treated vs. not-treated respondents (Tx). In our project, treatment refers to receipt of vaccination.
- 2) The survey data includes variables that characterize a respondent's demographic characteristics, such as age, gender, education, and health insurance.

We first developed a shell table of survey data, as presented below.

<i>Demographic characteristic</i>	<i>Sample size</i>	<i>Sample %</i>	<i>Weighted sample size</i>	<i>Weighted % (95% CI)</i>	<i>% Treated (95% CI)</i>
Total					
Age in years					
18-24					
25-35					
36-55					
56-65					
66+					
Gender					
Male					
Female					
Race/Ethnicity					
Non-Hispanic white					
Non-Hispanic black					
Hispanic					
Non-Hispanic other					
Education					
High school graduate or less					
Some college or Associate degree					
Bachelor's degree					
Master's degree or above					
Insurance*					
Private medical insurance					

<i>Demographic characteristic</i>	<i>Sample size</i>	<i>Sample %</i>	<i>Weighted sample size</i>	<i>Weighted % (95% CI)</i>	<i>% Treated (95% CI)</i>
Medicare					
Medicaid					
Other public insurance					
No Health Insurance					

CI: confidence interval

*Respondents may select more than one type of insurance

Note that the variables age, gender, race/ethnicity, and education represent single-select survey questions, in contrast to insurance, which represents a survey question for which respondents may select multiple responses.

“*Sample %*” is the column percent based on the unweighted sample size; it totals 100 for all categories of the variables that represent each single-select survey question. “*Weighted %*” is the column percent based on the weighted sample size; it also totals 100 for all categories of variables that represent each single-select survey question. “*Percent treated*” is the weighted percent of a given category’s population that is identified as treated (weighted row percent).

For a multiple choice survey item, such as insurance, the sums of “*Sample %*” and “*Weighted %*” may be greater than 100 because respondents may select more than one response. Thus, an individual respondent may be counted in more than one category of the variables representing insurance.

SAS MACROS TO DISPLAY RESULTS QUICKLY AND EFFICIENTLY

We then developed macros to display the results in the format outlined in the shell table.

MACRO 1: “TOTAL” ROW

The first macro – *TOTAL* – computes the first “Total” row of the table on a specified subset of the sample population. The macro call for %*TOTAL* is as follows:

```
%TOTAL(Tx, weight);
```

Tx refers to the treatment variable (in our example, receipt of vaccination), and *weight* refers to the weight variable. Macro %*TOTAL* is driven by two procedures, PROC SUMMARY AND PROC SURVEYFREQ.

PROC SUMMARY gives the sample size and weighted sample size:

```
PROC SUMMARY nway data=f noprint;
where &subset;
var N &weight;
output out = out(drop = _: ) sum = n wgt_n;
run;
```

PROC SURVEYFREQ calculates the total percent of the treated respondents and the lower and upper limits of the 95% confidence interval:

```
PROC SURVEYFREQ data=f nosummary;
where &subset;
tables &Tx/cl nostd;
ods output OneWay = total (keep = Tx Frequency WgtFreq Percent LowerCL UpperCL);
weight &weight;
run;
```

As expected, the sample percent and the weighted percent for the total row are 100. The macro %*TOTAL* results in the data set *total*, which carries all needed values for the first row in the table shell.

MACRO 2: COLUMN VALUES FOR SINGLE-SELECT SURVEY QUESTIONS

The second macro (%*SINGLE*) is intended to calculate the column values for single-select survey questions (such as age, gender, etc.) and allows the user to sort by either the column percent or the treatment percent, as well as subset the population for the table. The macro call is as follows:

```
%SINGLE (var, charact, fmt, Tx, weight, sort_column=, sort_cover=, subset=(0=0));
```

In this macro, *var* is a reported variable (age, for example), *charact* is the label that precedes the categories in the leftmost column of the shell (“Race/Ethnicity,” for example), and *fmt* is a user format with which the categories of the variable will be printed. *Tx* refers to the treatment variable, and *weight* refers to the weight variable. The two sorting

macro variables, *sort_column* and *sort_cover*, can be used to sort the variable categories in either ascending or descending order of the column percent (using *sort_column*) or treatment percent (using *sort_cover*). Sorting can be done in ascending order using an "A" and descending order using a "D." The *subset* function is used to indicate a where statement that will calculate the column values on a specified subset of the sample population. The default (0=0) indicates that the calculations will be done on the entire sample population.

The macro %*SINGLE* is used to obtain the necessary frequencies (column values). The core of the %*SINGLE* function contains two PROC SURVEYFREQs, one PROC FREQ, and a PROC SORT.

The first PROC SURVEYFREQ calculates the weighted percent of a given category's population that is identified as treated (weighted row percent), with a 95% confidence interval:

```
PROC SURVEYFREQ data=f nosummary;
  where &subset;
  tables &var*TX/cl row nostd;
  ods output CrossTabs = horiz;
  weight &weight;
run;
```

The data set *horiz* has all components (% Treated (95%CI)) of the estimates for all categories of the variable.

The second PROC SURVEYFREQ calculates the unweighted and weighted sample size for each category of the variable, as well as the weighted column percent and its 95% confidence interval:

```
PROC SURVEYFREQ data=f nosummary;
  where &subset;
  tables &var/cl row nostd;
  ods output Oneway = vertic(keep = &var frequency wgtFreq percent LowerCL UpperCL
  rename = (frequency = n wgtFreq = wgt_n));
  weight &weight;
run;
```

The data set *vertic* has all components (Sample size, Weighted sample size, Weighted % (95% CI)) of the estimates for all categories of the variable.

Next, PROC FREQ is used to calculate the unweighted percent for each category of the variable (unfortunately, PROC SURVEYFREQ does not calculate the unweighted percent), as follows:

```
PROC FREQ data=f;
  where &subset;
  tables &var/noprint out=unw (keep = &var percent rename = (percent = unw_pct));
run;
```

The data set *unw* has all components (sample column %) of the estimates for all categories of the variable.

Finally, we combine the separate data sets created thus far into one data set with the name of the variable, and use PROC SORT to sort the data set by either column percent or treatment coverage.

To illustrate this macro in practice, using the variables that represent single-select survey questions in the above table shell, the macro call is as follows:

```
%SINGLE (age, %NRBQUOTE (Age in years), agef, Tx, final_wgt, sort_column=A,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
%SINGLE (sex, %NRBQUOTE (Gender), sexf, Tx, final_wgt, sort_column=D,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
%SINGLE (race_ethn, %NRBQUOTE (Race/Ethnicity), racef, Tx, final_wgt, sort_cover=A,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
%SINGLE (education, %NRBQUOTE (Education), educationf, Tx, final_wgt, sort_cover=D,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
```

The first line of the macro call calculates the column values for age, sorted in ascending order of column percent, on the group of respondents who meet the condition BF1=2 or BF12=4. The second line is used to calculate the column values for sex, sorted in descending order of column percent, on the same subset. The third and fourth line calculates the values for race/ethnicity in ascending order of treatment coverage and education in descending order of treatment coverage, again on the same subset. In these macro calls, we use the %*NRBQUOTE* macro function to mask various symbols in the labels, such as " , ' , & , % , etc.

Each macro call ultimately creates a data set with the name of the variable the macro processes. Together, the data sets contain all of the numbers needed to fill the table shell. The following data step combines these data sets for printing all variables:

```
data combined_single;
set age sex race_ethn education;
```

MACRO 3: COLUMN VALUES FOR MULTIPLE CHOICE SURVEY QUESTIONS

The third macro (*%MULTY*) is intended to calculate the column values for each category of those survey questions for which respondents may select more than one response ("multiple choice" items, such as insurance). As a rule, multiple choice items in the SAS data set include several variables that represent the individual response items. For insurance (shown in the table shell) the variables are I1 – I5. Each variable can be selected (1) or not selected (0).

The macro call is as follows:

```
%MULTY (list_of_variables_in_group, list_of_labels, combined_ds, grouptitle, Tx,
weight, sort_column=, sort_cover=, subset= (0=0));
```

In this macro, *list_of_variables_in_group* is a list of the variables that represent the individual response options (for example, I1 I2 I3 I4 I5), *list_of_labels* is a list of the labels that precede the categories in the leftmost column of the shell (for example, "Private Medical Insurance"), and *combined_ds* is the name of the data set containing the numbers needed to fill the table shell. *Grouptitle* is the name of the characteristic (for example, "Insurance/No Insurance." *Sort_column* and *sort_cover* can be used to indicate that the individual response option rows should be sorted in ascending or descending order of column percent or treatment coverage percent, as described above. The subset function is used to indicate a where statement that will calculate the column values on a specified subset of the sample population, also as described above. Finally, *Tx* refers to the treatment variable, and *weight* refers to the weight variable.

At the core of the macro *%MULTY* are essentially the same PROC SURVEYFREQ and PROC FREQ as described above; however, the user should remember that contrary to *%SINGLE*, *%MULTY* only works with dichotomized variables (with values '1' and '0') and only the level '1' (selected) is the object of the estimate.

For the insurance multiple choice item in the above table shell the macro call is as follows:

```
%MULTY (I1 I2 I3 I4 I5,
%NRBQUOTE (Private medical insurance)
%NRBQUOTE (Medicare)
%NRBQUOTE (Medicaid)
%NRBQUOTE (Other public insurance)
%NRBQUOTE (No insurance),
insurance, %NRBQUOTE(Insurance/No Insurance),
Tx, final_wgt, sort_column=D, subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
```

The macro call ultimately creates a data set called *insurance* which contains all of the numbers needed to fill the table shell for that variable. As with the macro *%SINGLE*, we use PROC SORT to sort the data set by either column percent or treatment coverage.

RESULTS

To create a full table, we combine the datasets *total*, *combined_single*, and *insurance* and then print the dataset in the format of the table shell using PROC PRINT. The resulting table for the example described is presented below.

Demographic characteristic	Sample size	Sample %	Weighted sample size	Weighted % (95% CI)	% Treated (95% CI)
Total	3000	100	10464000	100	49.9 (47.1, 52.6)
Age					
66+	590	19.7	335000	3.2 (2.9, 3.5)	52.2 (48.2, 56.2)
56-65	284	9.5	1358000	13.0 (11.2, 14.8)	49.9 (42.6, 57.2)
36-55	1205	40.2	2253000	21.5 (19.8, 23.2)	50.4 (46.6, 54.1)
25-35	601	20	2893000	27.6 (25.3, 30.0)	51.3 (46.4, 56.2)
18-24	320	10.7	3625000	34.6 (31.7, 37.6)	48.2 (42.3, 54.0)
Gender					
Female	1513	50.4	5301515	50.7 (47.9, 53.4)	49.7 (45.9, 53.5)
Male	1487	49.6	5162485	49.3 (46.6, 52.1)	50.0 (46.1, 53.9)
Race/Ethnicity					
Hispanic	453	15.1	2234485	21.4 (19.0, 23.7)	48.0 (41.6, 54.3)
Non-Hispanic black	314	10.5	2476003	23.7 (21.0, 26.4)	48.4 (41.5, 55.3)

<i>Demographic characteristic</i>	<i>Sample size</i>	<i>Sample %</i>	<i>Weighted sample size</i>	<i>Weighted % (95% CI)</i>	<i>% Treated (95% CI)</i>
Non-Hispanic white	2082	69.4	5311387	50.8 (48.0, 53.5)	51.1 (47.8, 54.3)
Non-Hispanic other	151	5	442125	4.2 (3.2, 5.2)	52.7 (40.8, 64.6)
Education					
Master's degree or above	608	20.3	2089798	20.0 (17.8, 22.1)	53.3 (47.3, 59.3)
Some college or Associate degree	602	20.1	2073367	19.8 (17.6, 22.0)	53.2 (47.1, 59.4)
High school graduate or less	899	30	3194838	30.5 (28.0, 33.1)	49.7 (44.7, 54.8)
Bachelor's degree	891	29.7	3105997	29.7 (27.2, 32.2)	45.4 (40.5, 50.4)
Insurance*					
Private medical insurance	2090	69.7	7468136	71.4 (68.9, 73.8)	49.9(46.6,53.2)
Medicare	927	30.9	3203223	30.6 (28.1, 33.1)	50.0(45.2,54.9)
Other public insurance	612	20.4	2141661	20.5 (18.2, 22.7)	50.0(43.8,56.1)
Medicaid	589	19.6	2090449	20.0 (17.8, 22.2)	50.7(44.5,56.9)
No Health Insurance	410	13.7	1391595	13.3 (11.5, 15.1)	51.1(43.8,58.4)

CI: confidence interval

*Respondents may select more than one type of insurance

Note that the categories of each variable are sorted by either *Weighted %* or *% Treated*, as we specified in the macros. The sample size of 3000 refers to our subset that met the criteria BF1=2 or BF12 ne 4.

FLEXIBILITY

How flexible is our table? Suppose we need to replace *education* with *marital status* and place *marital status* after *insurance*. We would write the following statements:

```
%SINGLE (age, %NRBQUOTE (Age in years), agef, Tx, final_wgt, sort_column=A,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
%SINGLE (sex, %NRBQUOTE (Gender), sexf, Tx, final_wgt, sort_column=D,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
%SINGLE (race_ethn, %NRBQUOTE (Race/Ethnicity), racef, Tx, final_wgt, sort_cover=A,
subset=%NRBQUOTE (BF1=2 or BF12 ne 4));
/*%SINGLE (education, %NRBQUOTE (Education), educationf, Tx, final_wgt,
sort_cover=D, subset=%NRBQUOTE (BF1=2 or BF12 ne 4)); OLD LINE COMMENTED*/
%SINGLE (marital_status, %NRBQUOTE (Marital Status), maritalf, Tx, final_wgt,
sort_cover=D, subset=%NRBQUOTE (BF1=2 or BF12 ne 4)); /*NEW LINE*/
```

and then construct the data set for printing like this:

```
data forprint;
set total age sex race_ethn combined_multy marital_status;
```

where *combined_multy* is the combined insurance data created earlier. Could not be easier!

REFERENCE

1. Izrael, David; Ball, Sarah W.; and Donahue, Sara M.A. (2013). "Constructing a Table of Survey Data with Percent and Confidence Intervals in every Direction." *Proceedings of the NorthEast SAS Users Group Conference*.

CONTACT INFORMATION

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

Name: Rebecca V. Fink
E-mail: rebeccav_fink@abtassoc.com

All the numbers used in the table are based on randomly generated data and, therefore, have nothing in common with any survey data we have dealt with in our work.

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.