PROC SURVEYMEANS AND BAR GOERS IN SLO

Jasmin Cabarios, California Polytechnic University SLO, San Luis Obispo, CA

ABSTRACT

In the small town of San Luis Obispo, college students amass the respective bar scene, but do they really make up the majority of bar attendees? With a dataset and sample design that violates experimental design assumptions, we must use a tool that assumes we are violating said assumptions. We must think about how classical methods may not be appropriate for more complex, survey sample designs. We'll look at PROC SURVEYMEANS to give us some of those descriptive statistics that might have been misrepresented by more classic methods like PROC MEANS.

INTRODUCTION

Many a time have I found myself in Frog and Peach, a bar in San Luis Obispo (SLO), on a Thursday night. Usually, by my fifth or sixth beer, I would wonder about how many people go to the bars in SLO and not only that, I wanted to know the demographics of those who were in SLO bars.

I figured, I could find the average and confidence intervals to predict who participates in the bar scene in this college town. However, using regular methods like PROC MEANS wouldn't necessarily work with how we designed the survey above. When we have a sample that's surveyed and we want to subset a finite population, PROC MEANS more than likely will inflate confidence intervals and miscalculate degrees of freedom and errors. PROC SURVEYMEANS takes into account sub setting the population into strata and gives us an adjusted confidence interval that's more suited to our design.

PROC SURVEYMEANS OVERVIEW

Just as how I described above, normal means of data and statistical analysis will not work for surveyed data. Surveyed data doesn't follow normal assumptions for experimental statistical design. We have a finite population. We most likely won't meet the assumption of simple random sampling without replacement, especially so, since many people like to bar hop and these samples can be counted twice. This, in turn, will violate the independence assumption. We can also throw out identically distributed errors and the usual formulas for standard error of the mean. PROC SURVEYMEANS is the answer for this simple problem we face with clustered or stratified survey designs.

SYNTAX

This is the required statement to invoke the PROC SURVEYMEANS procedure:

```
PROC SURVEYMEANS; run;
```

This simple statement will take the most recent dataset in the SAS script and perform the procedure for all variables. That, in itself, won't give us the most meaningful descriptive statistics. Though, PROC SURVEYMEANS gives us really great options and is invoked very similar to the regular PROC MEANS procedure:

```
PROC SURVEYMEANS <options> <statistic-keywords>;
    BY variables;
    CLASS variables;
CLUSTER variables;
DOMAIN variables <variable*variable variable*variable*variable…> </option>;
POSTSTRATA variables / PSTOTAL = <option>;
RATIO <'label'> variables/variables;
REPWEIGHTS variables </ options>;
STRATA variables </ option>;
VAR variables;
WEIGHT variable;
```

PROC SURVEY MEANS invokes the procedure and we have options to name our datasets, specific descriptive statistics, and estimation methods. The BY statement is used to perform separate analyses for the different CLASS and VAR variables. VAR and CLASS both identifies variables to be analyzed as categorical variables; CLASS is used to identify numeric variables as categorical. With surveyed designs, sub setting a finite population is often done, we use the STRATA statement to identify the different subset or stratum. The WEIGHT statement identifies the sampling weight variable. DOMAIN identifies the variables that give us more in depth analysis for our subset, like comparing means.

Some of the more unusual options include POSTSTRATA, CLUSTER, RATIO, and REPWEIGHTS. POSTSTRATA identifies the variables for poststratification. CLUSTER will be used for cluster sample designs instead of STRATA which is used to stratified/subset sample designs. RATIO is used for ratio analytics for means or proportions. REPWEIGHTS identifies the weight variables used by other estimation methods like my favorite one to use for simulating replication, BOOTSTRAP. It is also important to note that all statements besides PROC SURVEYMEANS, WEIGHT, and POSTSTRATA can appear more than once.

PROC SURVEYMEANS = PROC MEANS

```
PROC SURVEYMEANS data = bar_goers;
    BY stratum;
    VAR total_goers;
run;
```

In this example, I am building upon the first one line PROC SURVEYMEANS procedure I invoked. I named my dataset, bar_goers My strata variable and the variable I want analyzed, total_goers. If we were to use the PROC MEANS statement in place of the PROC SURVEYMEANS and added the statistic options: MEAN, STDERR, CLM, the descriptive statistics for both PROCs would be identical.

WHEN PROC SURVEYMEANS ≠ PROC MEANS

```
PROC SURVEYMEANS data = bar_goers total=survey_misc;
    STRATA stratum / list;
    VAR total_goers;
    WEIGHT weight;
run;
```

I added three things from the example before. The total takes inflation factors, called the finite correlation factors, for each stratum into account. I defined the stratum with the STRATA statement and I added my sampling weight variable for each observation. PROC MEANS also has the WEIGHT statement but as we will explore further in this paper, PROC MEANS is not taking the inflation factors for each stratum and will inflate resulting confidence intervals.

ODS - SPECIFYING OPTIONS - PROC SURVEYMEANS

ODS GRAPH NAME	DESCRIPTION	PLOTS = OPTION	STATEMENT
BoxPlot	Box plots	BOXPLOT	PROC
DomainPlot	Box plots for domain stats for each domain definition	DOMAIN	DOMAIN
Histogram	Histograms with overlaid kernel densities and normal densities	HISTOGRAM	PROC
SummaryPanel	Histograms with overlaid kernel densities and normal densities, and box plots in a single panel	SUMMARY	PROC

Table 1. ODS Graphs Produced by PROC SURVEYMEANS (SAS INSTITUTE INC, 2019)

By default, PROC SURVEYMEANS will produce a summary panel and a histogram without specification. We can make boxplots through DOMAIN and comparing means and can also invoke other summary panels not in the initial PROC statement as indicated in Table1. Here is an example:

```
PROC SURVEYMEANS data=bar N=groups;
    STRATA date / list;
    VAR num_drinks;
    WEIGHT weight;
    ODS OUTPUT stratainfo = strata
               Statistics = bar_results;
run;
```

In this example, we have a simple dataset called bar, and we're identifying the sample size N= as the variable, groups. We also identified our STRATA , VAR and WEIGHT variables. Then, we specified our ODS OUTPUT. This is really cool, because summary tables for our data, stratum, and descriptive statistics are produced.

```
PROC SURVEYMEANS data=bar N=groups
                 mean clm sum clmsum t;
    STRATA date / list;
    VAR num_drinks;
    WEIGHT weight;
    ODS OUTPUT stratainfo = strata
               statistics = bar_results;
run;
```

In the PROC statement in this example, the descriptive statistics is specified to be included in the statistics table indicated by the ODS OUTPUT statement.

SIMPLE EXAMPLE OF CLASSICAL VS SURVEY MEAN METHODS

Before we get into more complex analytics and data manipulation for stratified survey samples, let's look at a simple stratified dataset called bar. The dataset has two variables, date (From August 10, 2020 to August 12, 2020) and the amount of drinks sold with alcohol content. Every observation represents a separate bar in SLO. To be able to do PROC SURVEYMEANS, a weight variable needs to be calculated and a separate data set of our stratum needs to be calculated. Let's say I have a record that on August 10, 2020 there were 300 groups, August 11 there were 275, and then on August 12, there were 310. Our weight is going to be the overall amount of groups for that day divided by the amount of observations (bars) that we sampled.

```
data bar;
    input date $7. num_drinks;

    if date = "10AUG20 " then weight = 300/10;
    if date = "11AUG20 " then weight = 275/10;
    if date = "12AUG20 " then weight = 310/10;

          datalines;
10AUG20 476
10AUG20 380
10AUG20 499
10AUG20 331     /*NOT REAL DATA WHO GOES TO A BAR DURING A PANDEMIC?!*/
10AUG20 310
10AUG20 254
10AUG20 427
10AUG20 218
10AUG20 496
10AUG20 378
11AUG20 203
11AUG20 357
11AUG20 463
11AUG20 269
11AUG20 240
11AUG20 285
11AUG20 325
11AUG20 421
11AUG20 286
11AUG20 379
12AUG20 281
12AUG20 425
12AUG20 493
12AUG20 227
12AUG20 458
12AUG20 484
12AUG20 382
12AUG20 391
12AUG20 283
12AUG20 211
;
run;
```

```
data groups;
    length date $7.;
    date = "10AUG20 ";
           _total_= 300;
           output;
date = "11AUG20 ";
    _total_ = 275;
    output;
date = "12AUG20 ";
    _total_ = 310;
    output;
run;

proc print data=bar;
title "Number of Drinks Sold at Bars in SLO: RAW SURVEY DATA";
run;
proc print data = groups;
title "Bar Strata";
run;
```

I would like to note that when you make a separate dataset with your strata, you have to name your total amount variable, _total_. There is no getting around it, since SAS looks for that variable to calculate the finite correlation factor and the main reason as to why PROC MEANS in lacking in computing descriptive statistics. Since we have done all the dataset manipulation that we needed, we can look at the classical and survey methods for calculating mean and 95% confidence interval.

```
proc surveymeans data=bar N=groups mean stderr clm;
    strata date / list;
    var num_drinks;
    weight weight;
    ods output stratainfo=strata
               statistics=bar_results;
title "PROC SURVEYMEANS";
run;

proc means data=bar mean stderr clm;
   var num_drinks;
    weight weight;
    by date;
title "PROC MEANS WEIGHTED";
run;

proc means data=bar mean stderr clm;
   var num_drinks;
   by date;
title "PROC MEANS NO WEIGHT";
run;
```

I have three methods of calculating mean, standard error of the mean, and 95% confidence intervals. These include. PROC SURVEYMEANS, PROC MEANS without weights, and PROC MEANS with weights. Note the similar syntax between PROC SURVEYMEANS AND PROC MEANS. Output 1. is shown below.

Statistics for the variable DIFF

Procedure	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
SURVEYMEANS	355.395480	17.433629	319.624627	391.166333
MEANS	354.4000000	17.4540843	318.7023895	390.097615
MEANS w/ weights	355.3954802	17.5170236	319.5691443	391.2218162

Output 1. Output from a CREATE TABLE Statement

Though the difference from method to method in this example isn't drastically different, the difference is still there. PROC MEANS without weights has the smallest interval, but has an inflated standard error of the mean and an inaccurate mean for our survey data. PROC SURVEYMEANS and PROC MEANS with weights have the same means as expected. However, PROC SURVEYMEANS has the most accurate standard error of the mean and a more precise interval than PROC MEANS with weights.

A MORE COMPLEX EXAMPLE WORKING WITH OTHER PROCS

In this example we are going to use a dataset called bar_goers. This dataset has three strata:

- California Polytechnic Students: C
- Not California Polytechnic Students below 35: N
- Not California Polytechnic Students above 35: O

We have 6 variables. The stratum variable is indicating which stratum each observation is made. The bar variable is indicating what bar the sample is observed. The groups variable is counting how many groups is flocking into a bar, in this case, each group is going to be counted as one as there are varying group size and to simplify our data. Alone is indicating every single person going into the respective bar, every pair is counted as two as we can easily record pairs, and all is all people in the bar.

```
data bar_goers;
    input stratum $1. bar :$15. groups alone pairs all;

    datalines;
C FrogAndPeach 29 39 52 172
C Library 78 96 67 308
C Motav 150 97 183 595
C McLintocks 100 97 183 545
C SideCar 100 30 78 286
C CreekyTiki 70 68 87 312
C TheGraduate 96 67 12 187
C BlackSheep 66 85 95 341
C BullsTavern 13 6 80 179
C Libertine 0 8 34 76
N FrogAndPeach 18 35 56 165
N Library 77 71 42 232
N Motav 20 9 37 103
N McLintocks 15 97 39 190
N SideCar 61 59 21 162
N CreekyTiki 13 4 44 105
N TheGraduate 98 46 84 312
N BlackSheep 95 87 17 216
N BullsTavern 25 54 58 195
N Libertine 5 1 38 82
O Library 13 0 12 37
O Motav 150 1 10 15 35
O CreekyTiki 5 2 6 19
O TheGraduate 7 3 8 26
O BlackSheep 4 11 9 33
O BullsTavern 3 13 6 28
O Libertine 10 11 14 49
;
run;
```

Just like the example comparing classical and survey methods, I need to make a dataset to indicate the total number of people in each stratum. I do that in the code below.

```
data survey_misc;        /* need to have a _total_ variable for PROC SURVEYMEANS */
    length stratum $1.;  /*_total_ is the total amount of people in each strata */
    input stratum $ _total_;
    datalines;
```

```
C 3001
N 1762
O 227
;
run;
```

At this point, we're still not done with setting up our dataset. If any case the data is not sorted, PROC SORT by the stratum variable is going to make analysis a lot easier. The dataset is already sorted by my stratum so we don't need to worry about that. The sampling weight is not calculated and in the classical vs survey example, the weight was calculated very rudimentary. We can use PROC MEANS and another data step to calculate weight and merge our output with PROC MEANS and preceding dataset.

```
proc means data = bar_goers;      /*makes another dataset for the frequency data */
    by stratum;
    var all;
    output out=mean_data n=n;
run;

data bar_goers;                /* merging all the datasets together by our strata */
    merge bar_goers survey_misc mean_data;
    by stratum;
    weight = _total_/n;       /*calculating the sampling weight for each obs */
    drop _TYPE_;              /* Dropping the variable _TYPE_ since all type = 0 */
run;

proc print data = bar_goers;
title "Bar Scene: Simulated Raw Data";
run;
```

Bar Scene: Simulated Raw Data

Obs	stratum	bar	groups	alone	pairs	all	_total_	_TYPE_	_FREQ_	n	weight
1	C	FrogAndPeach	29	39	52	172	3001	0	10	10	300.100
2	C	Library	78	96	67	308	3001	0	10	10	300.100
3	C	Motav	150	97	183	595	3001	0	10	10	300.100
4	C	McLintocks	100	97	183	545	3001	0	10	10	300.100
5	C	SideCar	100	30	78	286	3001	0	10	10	300.100
6	C	CreekyTiki	70	68	87	312	3001	0	10	10	300.100
7	C	TheGraduate	96	67	12	187	3001	0	10	10	300.100
8	C	BlackSheep	66	85	95	341	3001	0	10	10	300.100
9	C	BullsTavern	13	6	80	179	3001	0	10	10	300.100
10	C	Libertine	0	8	34	76	3001	0	10	10	300.100
11	N	FrogAndPeach	18	35	56	165	1762	0	10	10	176.200
12	N	Library	77	71	42	232	1762	0	10	10	176.200
13	N	Motav	20	9	37	103	1762	0	10	10	176.200
14	N	McLintocks	15	97	39	190	1762	0	10	10	176.200
15	N	SideCar	61	59	21	162	1762	0	10	10	176.200
16	N	CreekyTiki	13	4	44	105	1762	0	10	10	176.200
17	N	TheGraduate	98	46	84	312	1762	0	10	10	176.200
18	N	BlackSheep	95	87	17	216	1762	0	10	10	176.200
19	N	BullsTavern	25	54	58	195	1762	0	10	10	176.200
20	N	Libertine	5	1	38	82	1762	0	10	10	176.200
21	O	Library	13	0	12	37	227	0	7	7	32.429
22	O	Motav	150	1	10	15	227	0	7	7	32.429
23	O	CreekyTiki	5	2	6	19	227	0	7	7	32.429
24	O	TheGraduate	7	3	8	26	227	0	7	7	32.429
25	O	BlackSheep	4	11	9	33	227	0	7	7	32.429
26	O	BullsTavern	3	13	6	28	227	0	7	7	32.429
27	O	Libertine	10	11	14	49	227	0	7	7	32.429

Display 1. Output of final updated dataset.

Note the variables calculated in the PROC MEANS procedure and preceding data step to create the survey_misc dataset , _total_, _FREQ_, and n. We used _total_ divided by n to calculate our sampling weight. Now that we have all the variables needed to do a PROC SURVEYMEANS calculation, let's go ahead and do that.

```
proc surveymeans data = bar_goers
                 total=survey_misc /*total= finite correlation factor */
                 sum clsum mean clm t;
     strata stratum / list;
  /* can include BY STRATUM statement to have separate analysis for each stratum */
     var all;
     weight weight;
     ods output stratainfo=strata
                statistics=bar_goers_results;
  run;
```

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	3
Number of Observations	27
Sum of Weights	4990

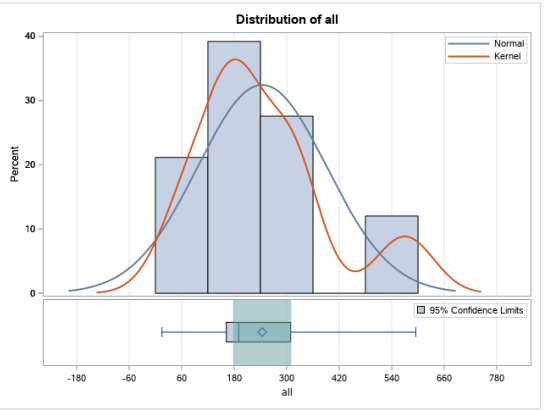
Stratum Information

Stratum Index	stratum	Pop. Total	Sampling Rate	N Obs	Variable	N
1	C	3001	0.33%	10	all	10
2	N	1762	0.57%	10	all	10
3	O	227	3.08%	7	all	7

Statistics

Variable	Mean	Std Error of Mean	t Value	Pr > |t|	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
all	244.043530	32.146793	7.59	<.0001	177.695810	310.391249	1217777	160412	886702.094	1548852.33

The SURVEYMEANS Procedure



Display 2. Output of PROC SURVEYMEANS for the bar_goers dataset

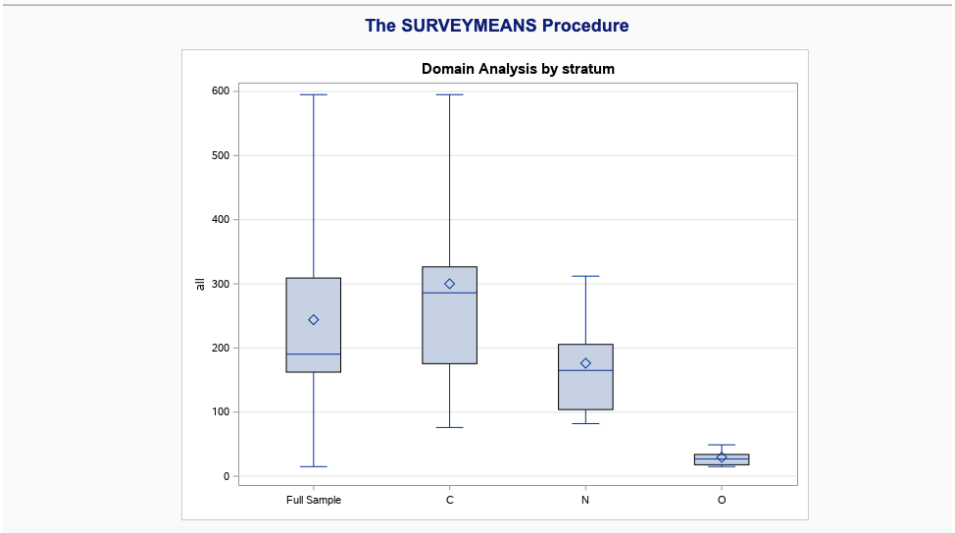
If we look at Display 2. We will see the output for PROC SURVEYMEANS with the bar_goers dataset. In the PROC statement we have a statement TOTAL=, this works similar to the N= statement in the classical vs survey method example. TOTAL= is computing a correlation factor that is to adjust inflation with finite population within each strata. There are three summary tables, the overall totals in the Data Summary table, all the stratum information and the descriptive statistics. The output includes all the statistics specified in the SAS script. This includes the t test and the p value. In this context, the t test is testing whether or not the mean is 0. The pvalue is incredibly small so we conclude that the mean is not zero. An inferential finding like that can leave more to be desired, but with the DOMAIN statement we can compare difference of means and have a really cool side-by-side boxplot that visualize the prospective mean difference.

```
proc surveymeans data = bar_goers total=survey_misc
                 sum clsum mean clm t;
     strata stratum / list;
```

```
    var all;
    weight weight;
    domain stratum / CLDIFF; /*bolded to show the difference in the example above*/
    ods output stratainfo=strata
               statistics=bar_goers_results;
run;
```

Differences of all Means for stratum Domains

stratum	-stratum	Diff Estimate	Std Error	DF	t Value	Pr > |t|	95% Confidence Limits	
C	N	123.900000	56.308733	9	2.20	0.0553	-3.479205	251.279205
C	O	270.528571	52.060053	9	5.20	0.0006	152.760550	388.296593
N	O	146.628571	22.284372	9	6.58	0.0001	96.217820	197.039323



Display 3. Modification of the PROC SURVEYMEANS Output with DOMAIN and CLDIFF

The table that describes the differences of all the means for the strata is similar to a Tukey's comparison of means. Here we see that at a 0.05 significance level, we have strong statistical evidence that there are two significant differences with demographics in SLO bars. Cal Poly students and non Cal Poly students over 35 and with Non Cal Poly students younger than 35 and non Cal Poly students over 35. There isn't enough statistical difference in the demographics in SLO bars between Cal Poly students and non Cal Poly students younger than 35.

CONCLUSION

Classical methods like PROC MEANS is great for it's intended use, simple random samples and experimental designs. We should take into consideration whether our design is indeed experimental, surveyed, clustered, observational or other; because if we use non-survey tools for analysis for surveyed data, we will get miscalculated descriptive statistics.

PROC MEANS and PROC SURVEYMEANS are very similar in regards with syntax, we can even get the same outputs if we manipulate PROC SURVEYMEANS correctly. The difference, however, lies in the TOTAL= statement. Without the computation of the finite population correlation factor within each strata being taken into account, PROC SURVEYMEANS will be lacking in computing correct standard error and confidence intervals.

Thus, we can conclude that using the wrong PROC for analytics can come with misrepresented data and conclusions. Because of the fact that PROC MEANS and PROC SURVEYMEANS have very similar syntax, there really isn't a reason to use the wrong PROC if we have experimental data or surveyed data.

REFERENCES

An, A., Watts, D. "New SAS Procedures for Analysis of Sample Survey Data." SAS Institute, Inc. Available at https://stats.idre.ucla.edu/wp-content/uploads/2016/02/svy_survey.pdf

"Applied survey data analysis using SAS 9.4." UC REGENTS. Available at https://stats.idre.ucla.edu/sas/seminars/sas-survey/.

"The SURVEYMEANS procedure." SAS Institute, Inc. 13 Dec, 2019. Available at https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_surveymeans_syntax05.htm&docsetVersion=15.1&locale=en.

ACKNOWLEDGMENTS

All datasets and analysis are fictional and does not have merit in regards to the actual demographics in SLO bars.

RECOMMENDED READING

- Stratified Random Sample: Definition, Examples https://www.statisticshowto.com/stratified-random-sample/

CONTACT INFORMATION

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

Jasmin Cabarios
California Polytechnic University San Luis Obispo
XXX CIRCLE
Stockton, CA, 95209
(209) 603-5503
jcabario@calpoly.edu