Live Session 09

Week	Date	Plan
09	03/15	Complex designs
10	03/22	Ratio estimation
11	03/29	Categorical data analysis
12	04/05	Project Presentations
13	04/12	Review/Lab 13
14	04/18	Final Exam

Lab 8

- The exercise is to talk about what happens when the cluster and strata statements are in or out
- Missing weight statement
- wt= M/m
- wt=11426/64

Example (Cluster Design)

- **Source :** Analyzes the data in Example 5.6 of Sampling: Design and Analysis, 2nd ed. by S. Lohr. Copyright 2008 by Sharon Lohr

- Cluster designs are often used in educational studies, since students are naturally clustered in to classrooms or schools. Consider a population of 187 high school algebra classes in a city. An investigator takes an SRS of 12 of those classes and give each student in the sampled classes a test about function knowledge. The data are given in the file algebra.csv

- Weights for each observation = 187/12

Example (Cluster Design)

- **Source :** Analyzes the data in Example 5.6 of Sampling: Design and Analysis, 2nd ed. by S. Lohr. Copyright 2008 by Sharon Lohr

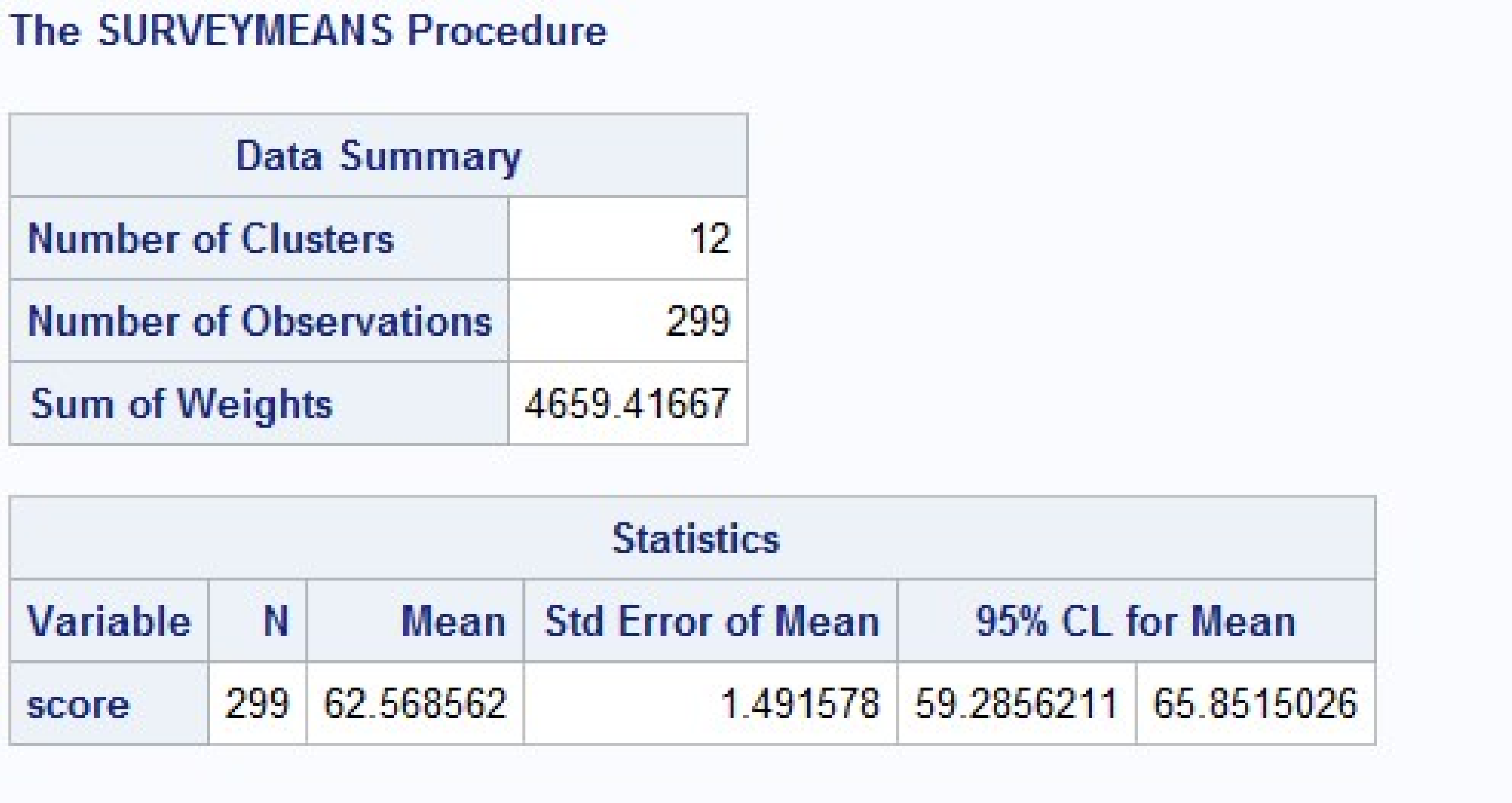
- Estimate the mean score in the population!

```
filename algebra
'C:\Users\Mahesh\Desktop\algebra.csv';

data algebra;
   infile algebra delimiter= ',' firstobs = 2;
   input class Ni score;
   sampwt = 187/12;
run;



proc surveymeans data=algebra total = 187 ;
   cluster class;
   var score;
   weight sampwt;
run;
```



The SURVEYMEANS Procedure

Data Summary	
Number of Clusters	12
Number of Observations	299
Sum of Weights	4659.41667

Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
score	299	62.568562	1.491578	59.2856211	65.8515026

Example (Two-stage Design)

- **Source :** Analyzes the data in Exercise 5.6 of Sampling: Design and Analysis, 2nd ed. by S. Lohr. Copyright 2008 by Sharon Lohr

- An inspector samples cans from a truckload of canned creamed corn to estimate the total number of worm fragments in the truckload. The truck has 580 cases; each case contains 24 cans. The inspector samples 12 cases at random, and subsample 3 cans randomly from each selected case.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Can1	1	4	0	3	4	0	5	3	7	3	4	0
Can2	5	2	1	6	9	7	5	0	3	1	7	0
Can3	7	4	2	6	8	3	1	2	5	4	9	0

```
data worms;
do case = 1 to 12;
do can = 1 to 3;
input worms @@;
wt = (580/12)*(24/3);
output;
end;
end;
cards;
1  5  7
4  2  4
0  1  2
3  6  6
4  9  8
0  7  3
5  5  1
3  0  2
7  3  5
3  1  4
4  7  9
0  0  0
;
```

```
proc surveymeans data=worms total = 580 sum
clsum;
weight wt;
cluster case;
var worms;
run;
```

Estimation from two-stage designs

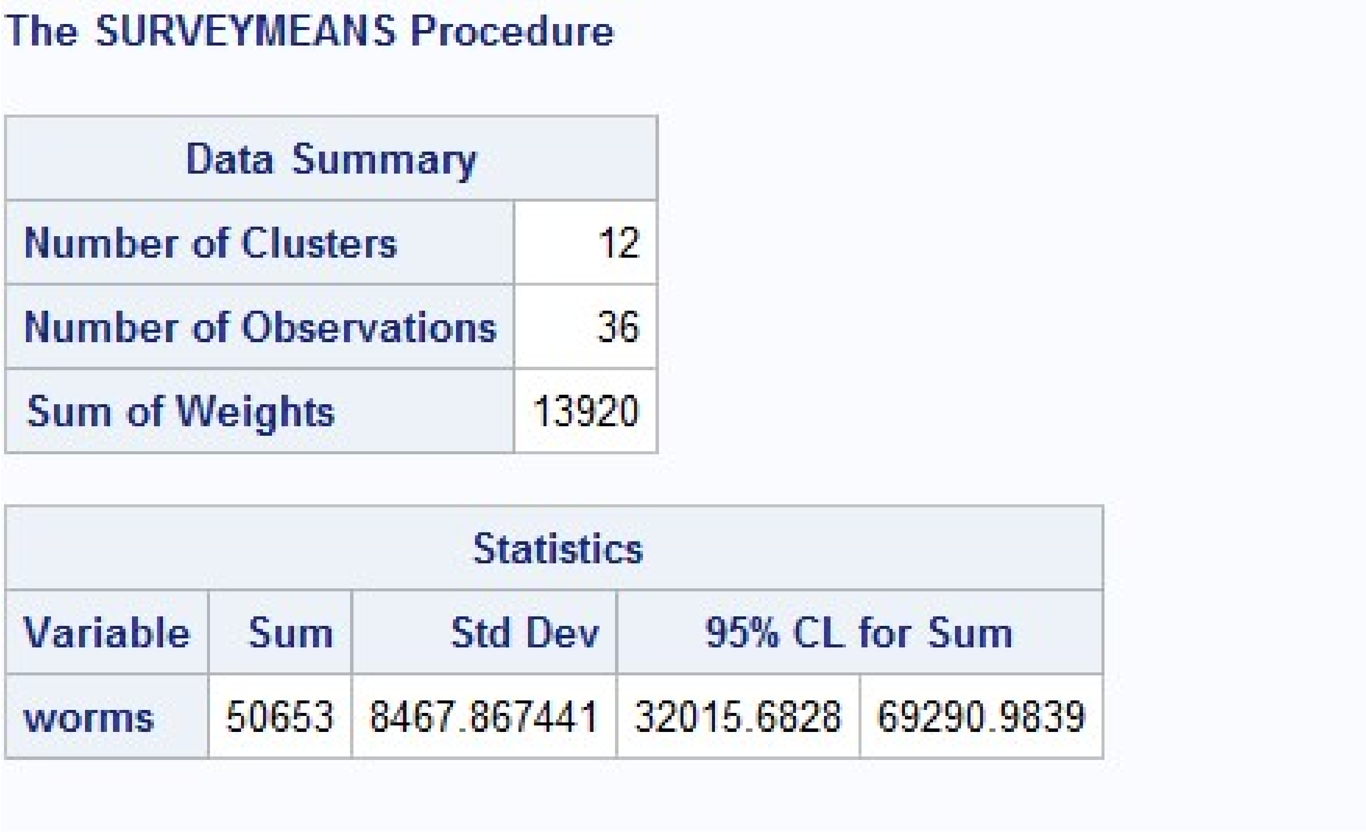
(with no stratification)

- Suppose there are N_i SSU's in the i^{th} PSU, and M PSU's in the population. If you sample m PSU's and n_i of the N_i SSU's, then the probability of selection is

$$\frac{m}{M}\frac{n_i}{N_i}$$

- Then the weight for the jth SSU in the ith PSU is

$$w_{ij} = \frac{M}{m}\frac{N_i}{n_i}$$



The SURVEYMEANS Procedure

Data Summary	
Number of Clusters	12
Number of Observations	36
Sum of Weights	13920

Statistics				
Variable	Sum	Std Dev	95% CL for Sum	
worms	50653	8467.867441	32015.6828	69290.9839

Estimation from two-stage designs

(with no stratification)

- Suppose there are N_i SSU's in the i^{th} PSU, and M PSU's in the population. If you sample m PSU's and n_i of the N_i SSU's, then the probability of selection is

$$\frac{m}{M}\frac{n_i}{N_i}$$

- Then the weight for the jth SSU in the ith PSU is

$$w_{ij} = \frac{M}{m}\frac{N_i}{n_i}$$

Estimation from two-stage designs

(with stratification)

- Suppose there are N_{hi} SSU's in the i^{th} PSU, and M_h PSU's in the h^{th} stratum. If you sample m_h PSU's and n_{hi} of the N_{hi} SSU's, then the probability of selection is

$$\frac{m_h}{M_h}\frac{n_{hi}}{N_{hi}}.$$

and the weights are $w_{ij} = \dfrac{M_h}{m_h}\dfrac{N_{hi}}{n_{hi}}$

Data Analysis for two-stage designs in SAS

- PROC SURVEYMEANS can be used for analysis of two or more-than-two-stage (called multi-stage) designs
  - You must specify the weights
  - If there are strata, they must be specified as usual
  - You must specify the primary sampling unit identifier
- Even if there are more than two stages, you only need to specify the PSU's, though weights based on all stages must be incorporated.

Email Task

A large manufacturing company has policies against employees using their email system for certain purposes. For example:

non-work communication,

ones discussing certain types of proprietary information, etc.).

Each email can be classified into one of 4 categories:

a non-violation, or one of 3 categories of violation.

Though the company has automated monitoring in place, they would like to supplement this for accuracy by sampling the emails and having a human inspector.

Going forward, they will produce estimates of the proportion of emails in each categories each month based on the sample results.

Categories: a, b, c, d

This will allow them to both monitor the policy violation rate over time, and to compare results with their automated monitoring system.

Email Task

- Suppose that the company realizes that there seem to be more policy violations on Fridays than other days of the week.
- They decide to select two Fridays each month at random, and two non-Fridays each month at random, and select 50 emails from each day.
- What kind of design is this?
  - stratified, two-stage design
- What is probability of selection?
  - For Fridays
    - (2/# of Fridays in month)*(50/#of emails in selected day)
  - For other days
    - (2/# of non-Fridays in month)*(50/#of emails in selected day)
- What is weight?
  - reciprocals of above

Data Table

Stratumid Fri or non-Fri)	Psuid (days)	Ssuid (emails)	Cat (a,b,c, or d)	wt1	wt2	base wt
1	1	1	a			
…	…	…	…			
1	1	50	b			
1	2	1	a			
…	…	…	…			
1	2	50	c			
2	1	1	a			
…	…	…	…			
2	1	50	a			
2	2	1	a			
…	…	…	…			
2	2	50	d			

Probabilities of selection for Fridays

Suppose 4 Fridays

Prob of selecting a Fri

2/4 = 1/2
wt1 = 2

1st selected Friday has 200 emails

prob of selecting an email =

50/200 = 1/4
wt2 = 4
basewt = 2*4=8

2nd selected Friday has 250 emails

prob of selecting an email =

50/250 = 1/5
wt2 = 5
basewt = 2*5 = 10

Data Table with weights for Fri

Stratumid Fri or non-Fri)	Psuid (days)	Ssuid (emails)	Cat (a,b,c, or d)	wt1	wt2	base wt
1	1	1	a	2	4	8
…	…	…	…	…	…	…
1	1	50	b	2	4	8
1	2	1	a	2	5	10
…	…	…	…	…	…	…
1	2	50	c	2	5	10
2	1	1	a			
…	…	…	…			
2	1	50	a			
2	2	1	a			
…	…	…	…			
2	2	50	d			

Probabilities of selection for non-Fri

Suppose 16 non-Fridays

Prob of selecting a non-Fri

2/16 = 1/8

wt1 = 8

1st selected non-Friday has 400 emails

prob of selecting an email =

50/400 = 1/8

wt2 = 8

basewt = 8*8=64

2nd selected non-Friday has 500 emails

prob of selecting an email =

50/500 = 1/10

wt2 = 10

basewt = 8*10 = 80

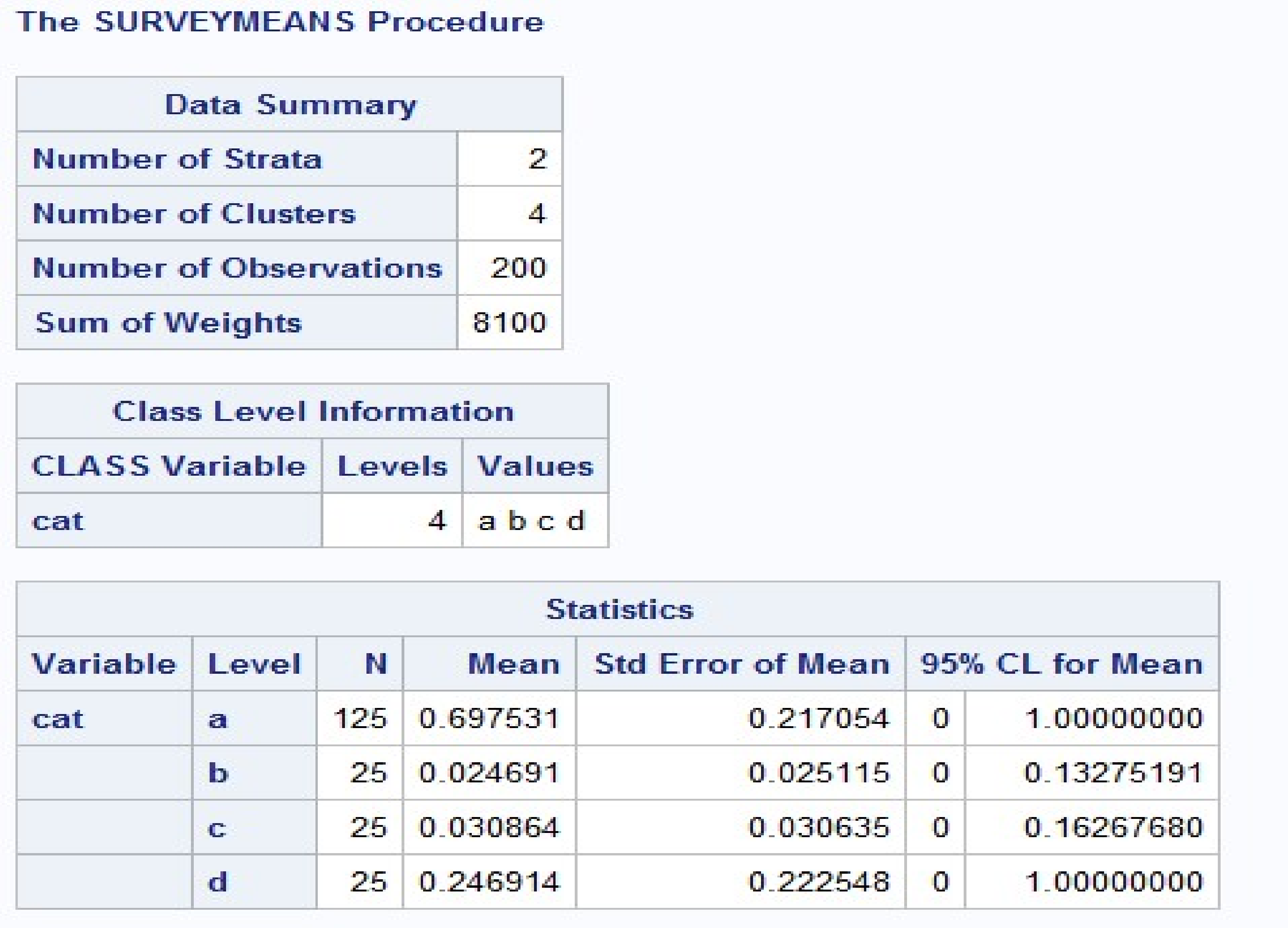
Data Table with Fri & non-Fri weights

Stratumid Fri or non-Fri)	Psuid (days)	Ssuid (emails)	Cat (a,b,c, or d)	wt1	wt2	base wt
1	1	1	a	2	4	8
…	…	…	…	…	…	…
1	1	50	b	2	4	8
1	2	1	a	2	5	10
…	…	…	…	…	…	…
1	2	50	c	2	5	10
2	1	1	a	8	8	64
…	…	…	…	…	…	…
2	1	50	a	8	8	64
2	2	1	a	8	10	80
…	…	…	…	…	…	…
2	2	50	d	8	10	80

SAS code for estimates (without fpc)

```
proc surveymeans data = email ;
title 'analysis of stratified cluster design without fpc';
class cat;
strata stratumid;
cluster psuid;
weight basewt;
var cat;
run;
```

Proc Survey means without fpc



The SURVEYMEANS Procedure

Data Summary	
Number of Strata	2
Number of Clusters	4
Number of Observations	200
Sum of Weights	8100

Class Level Information		
CLASS Variable	Levels	Values
cat	4	a b c d

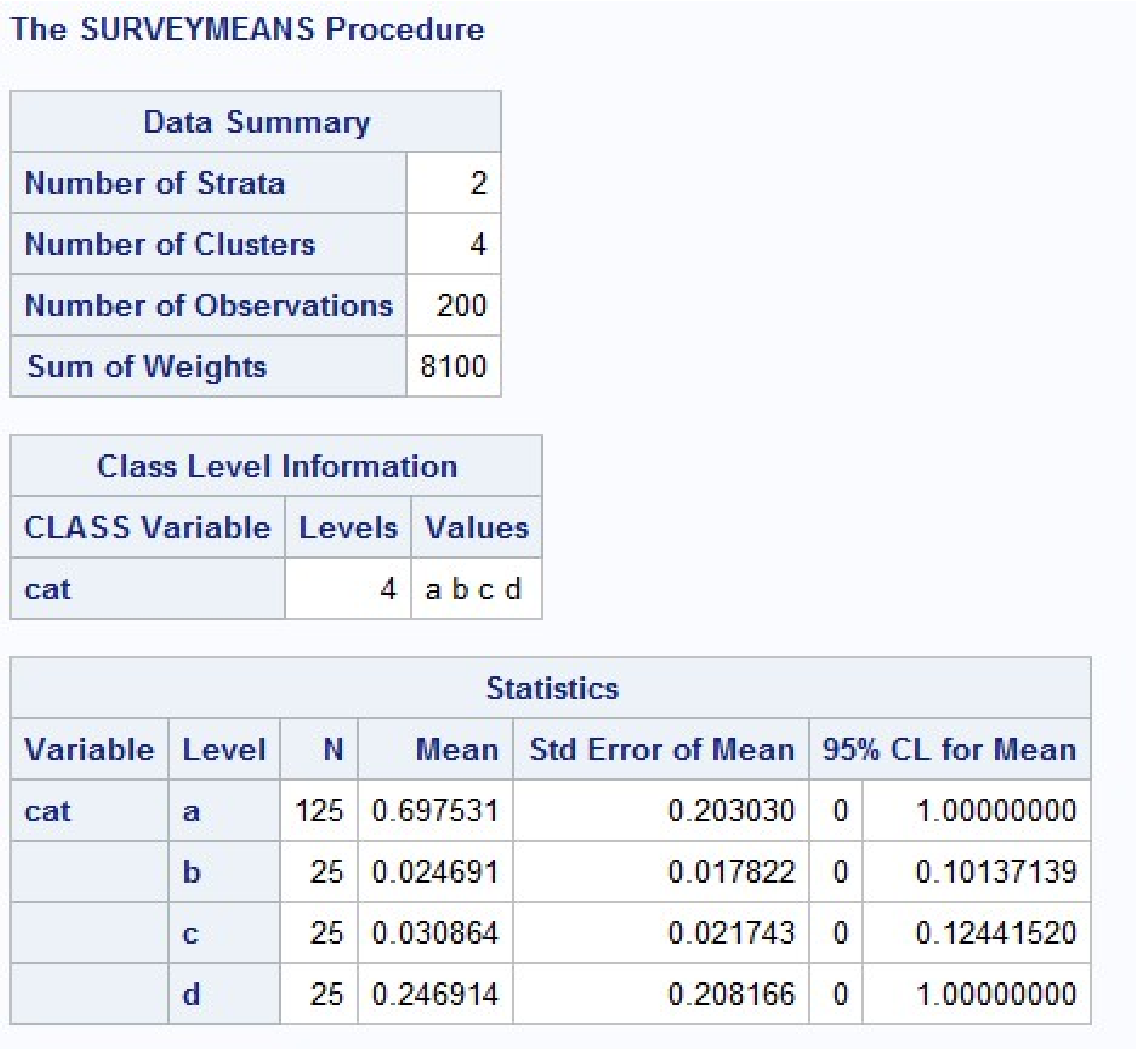
Statistics						
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
cat	a	125	0.697531	0.217054	0	1.00000000
	b	25	0.024691	0.025115	0	0.13275191
	c	25	0.030864	0.030635	0	0.16267680
	d	25	0.246914	0.222548	0	1.00000000

Large standard errors without fpc

SAS code for estimates (with fpc)

```
data strsizes;
stratumid = 1;  _total_= 4;   output;
stratumid = 2; _total_ = 16; output;
;
proc surveymeans data = email total = strsizes;
title 'analysis of stratified cluster design with fpc';
class cat;
strata stratumid;
cluster psuid;
weight basewt;
var cat;
run;
```

Proc Survey means with fpc



The SURVEYMEANS Procedure

Data Summary	
Number of Strata	2
Number of Clusters	4
Number of Observations	200
Sum of Weights	8100

Class Level Information		
CLASS Variable	Levels	Values
cat	4	a b c d

Statistics

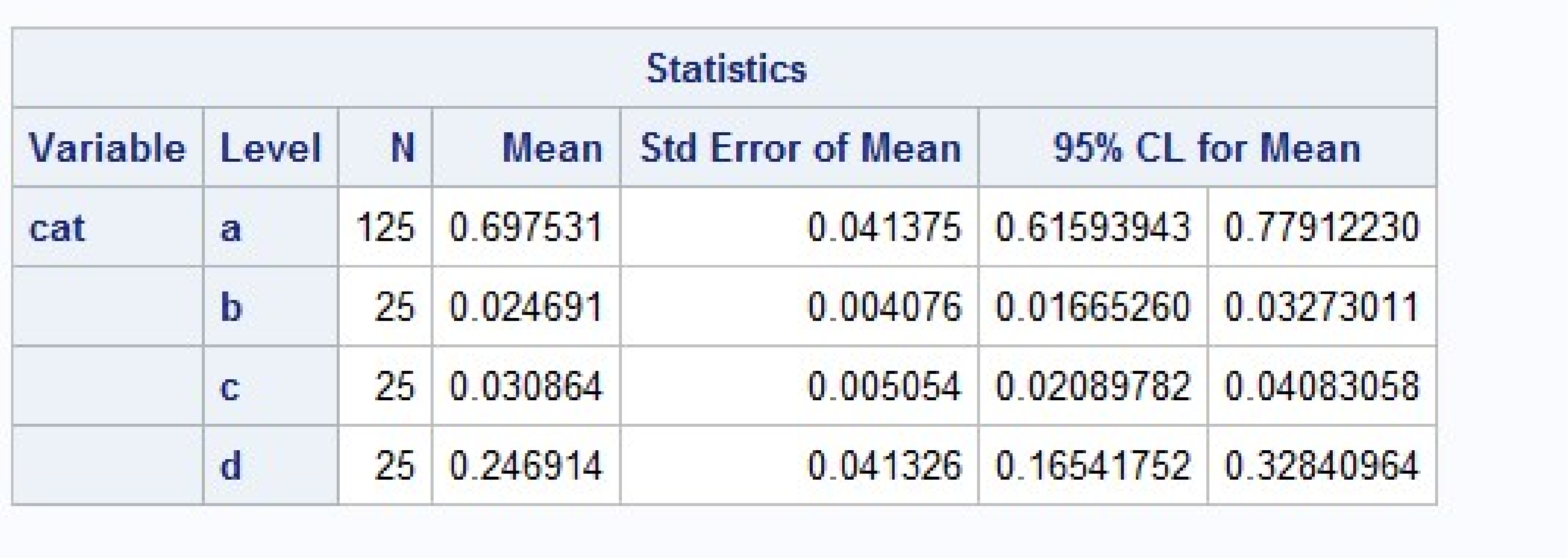
Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
cat	a	125	0.697531	0.203030	0	1.00000000
	b	25	0.024691	0.017822	0	0.10137139
	c	25	0.030864	0.021743	0	0.12441520
	d	25	0.246914	0.208166	0	1.00000000

Everything the same except confidence intervals.

SAS code not accounting for clustering

```
data strsizes;
stratumid = 1;  _total_=900;   output;
stratumid = 2; _total_ = 7200; output;
;
proc surveymeans data = email total = strsizes;
title 'analysis as if it was NOT a two-stage design';
strata stratumid;
weight basewt;
var cat;
run;
```

Proc Surveymeans without clustering
(as if not a 2-stage design)



Statistics

Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean	
cat	a	125	0.697531	0.041375	0.61593943	0.77912230
	b	25	0.024691	0.004076	0.01665260	0.03273011
	c	25	0.030864	0.005054	0.02089782	0.04083058
	d	25	0.246914	0.041326	0.16541752	0.32840964

Great confidence intervals! BUT they are not true --- don’t fool yourself