The Use of One and Two Stage Cluster Sampling with Probabilities Proportional to Size to Estimate the Total Number and Proportion of Single Family Units in a Neighborhood

A Case Study of the South Tabor Neighborhood in Portland, Oregon

Marie L. Tree

STAT422

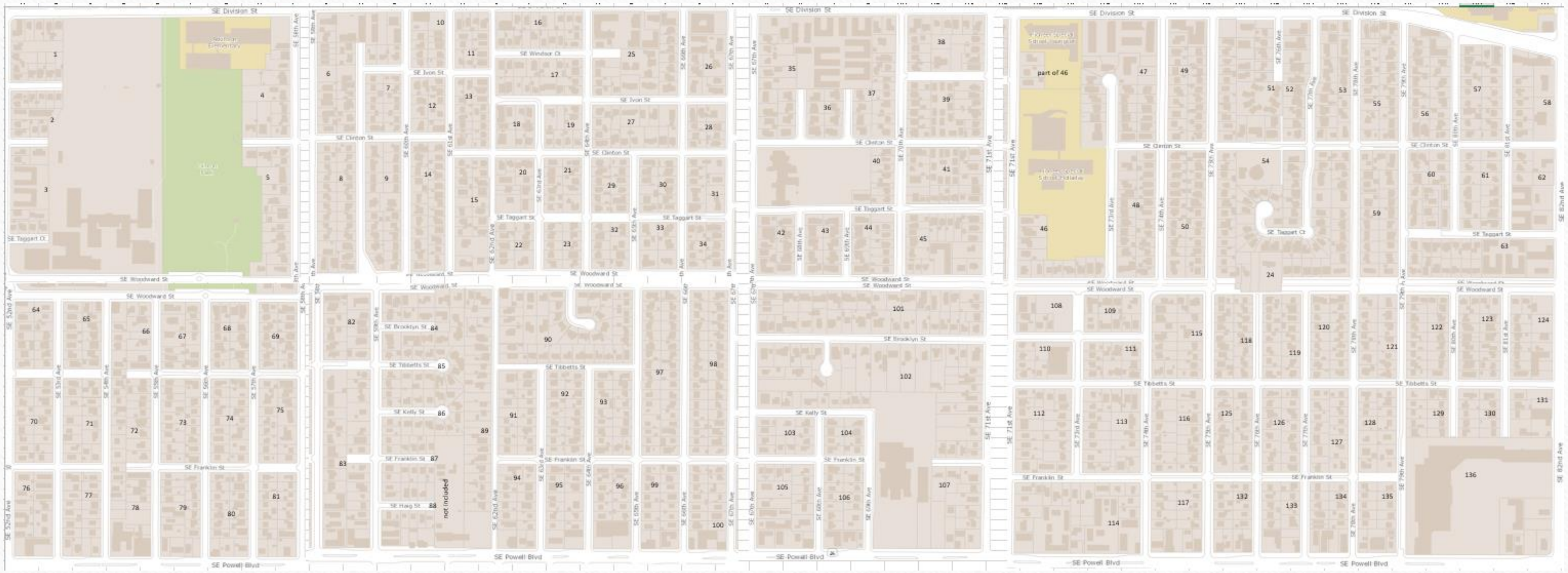
Introduction

- What is the research question? The question of this project is "what is an effective sample survey method to obtain an estimation of the total number and proportion of single housing units in a neighborhood?"
- What is the population to be sampled?: A case study of the surveyor's neighborhood---the South Tabor neighborhood of Portland, Oregon--- will be conducted.
- What will be measured?: Areas designated as plots by the City of Portland will be measured with 1/0 responses depending on whether or not a single family unit is on the plot.
- Why is the sample being taken?: The surveyor is passionate about Portland, Oregon and her neighborhood, and has a personal interest in learning more about where she lives.

Sampling Design

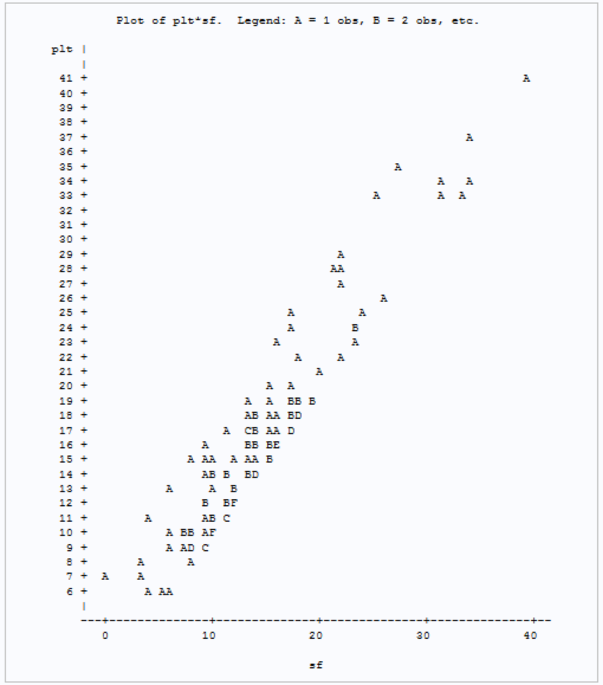
- What are the elements?: The elements are plots of land with distinct boundaries as marked by the City of Portland.
- What is the population?: The population is the collection of plots that make up the South Tabor neighborhood. The area that constitutes the population is bordered by SE Division Street to the north, SE 82nd Street to the east, SE Powell Boulevard to the south, and SE 52nd Street to the west.
- What are the sampling units?: The sampling units are blocks. The blocks may be a regularly or irregularly shaped area with an identifiable geographical boundary. The blocks were numbered by the surveyor in the frame.
- What is the frame?: The frame was obtained from the website portlandmaps.com.

The Frame of the South Tabor Neighborhood in Portland, Oregon



Census Data and One Stage Cluster Sample Size Determination for Probabilities Proportional to Size

y_i versus m_i for the population



Total Plots	Total Single Family Units	Proportion of Single Units	δ used in sample size determinatio n
2209	1919	.8687	2.23

$n=(N\delta^2)/((ND)+\delta^2)$

Let B=100

$D=B^2/4N^2=100^2/(4(136^2))=.135$

$n=(136*2.23^2)/((136*.135)+2.23^2)=28.985 \uparrow 30$

- The macro pps was used with the seed=40567 to generate a random sample of blocks from the census data

Single Stage Cluster Sampling Using Probabilities Proportional to Size Results

	block	plots	single fam	yibar	(yibar-uhatpps)**2
1	5	16	15	0.9375	0.00375769
2	6	29	22	0.7586206897	0.0138248942
3	7	12	12	1	0.01532644
4	20	10	8	0.8	0.00580644
5	43	10	10	1	0.01532644
6	47	25	17	0.68	0.03849444
7	47	25	17	0.68	0.03849444
8	49	21	20	0.9523809524	0.0058035375
9	51	28	21	0.75	0.01592644
10	54	33	31	0.9393939394	0.003993474
11	54	33	31	0.9393939394	0.003993474
12	57	18	17	0.9444444444	0.0046573042
13	59	22	22	1	0.01532644
14	68	14	14	1	0.01532644
15	83	28	22	0.7857142857	0.0081876645
16	83	28	22	0.7857142857	0.0081876645
17	83	28	22	0.7857142857	0.0081876645
18	83	28	22	0.7857142857	0.0081876645
19	84	11	9	0.8181818182	0.0033661094
20	90	41	39	0.9512195122	0.0056279272
21	90	41	39	0.9512195122	0.0056279272
22	93	18	18	1	0.01532644
23	93	18	18	1	0.01532644
24	95	15	13	0.8666666667	0.0000908844
25	98	34	34	1	0.01532644
26	100	17	13	0.7647058824	0.0124309383
27	114	23	16	0.6956521739	0.0325975175
28	116	16	16	1	0.01532644
29	118	17	17	1	0.01532644
30	135	14	10	0.7142857143	0.0262162359

- $\mu_{(hat)pps}=(1/n)\sum\bar{y}_i$

$=(1/30)[0.9375+0.7586+\ldots0.71429]=0.8762$

- $V_{(hat)}\,(u_{(hat)pps})=(1/(n(n-1)))\sum(\bar{y}_i-\mu_{(hat)pps})^2$

$=(1/(30(29)))[0.00375+0.01382+\ldots+0.02622]$--->B=0.042

- $\tilde{\iota}_{(hat)pps}=(M/n)\sum\bar{y}_i$

$=(2209/30)[0.9375+0.7586+\ldots0.71429]=1935.564$

- $V_{(hat)}\,(\tilde{\iota}_{(hat)pps})=(M^2/(n(n-1))\sum(\bar{y}i-\mu_{(hat)pps})^2$

$=(2209^2/(30*29))[0.00375+0.01382+\ldots+0.02622]$--->B=93.707

- See appendix for SAS code and results

Two Stage Cluster Sample Size Determination for Probabilities Proportional to Size

- Pilot Study

Simple Random Sample
Output Data Set = project_srs

Obs	blck	plt	sf
1	9	23	23
2	14	24	23
3	29	12	12
4	35	14	13
5	43	10	10
6	57	18	17
7	60	15	8
8	82	12	11
9	99	17	14
10	111	10	10

SAS Results for

δ_w^2 and δ_b^2

Covariance Parameter Estimates	
Cov Parm	Estimate
blck	543.41
Residual	23.8534

- $m=\sqrt{(\delta_w^2/\delta_b^2)}=\sqrt{(23.8534/543.41)}=.2095\uparrow 21\%$

  Therefore, 21% within clusters will be sampled
- The desire is to use the same random sample of blocks that was used in One Stage Clustering
- n=30, or 22% of the total clusters

  $(V_{(hat)}(\mu_{(hat)}))=(1/n)(\delta_b^2+(\delta_w^2/m))$

  $=(1/.22)(543.41+(23.8534/.21))=2986.35$

  --->B=110

Two stage Cluster Sampling Using Probabilities Proportional to Size Results

	block	plots	single fam	samplesize	sample size Rounded	Plots Sampled	#,single fam	y(bar)i	y(bar)i-u(hat)pps
1	5	16	15	3.36	4	[2, 5, 8, 14].	4	1	0.1315
2	6	29	22	6.09	7	[1, 4, 5, 14, 17, 26, 29]	4	0.5714285714	-0.297071429
3	7	12	12	2.52	3	[2, 3, 8]	3	1	0.1315
4	20	10	8	2.1	3	[1, 6, 8]	2	0.6666666667	-0.201833333
5	43	10	10	2.1	3	[1, 9, 10]	3	1	0.1315
6	47	25	17	5.25	6	[1, 5, 11, 15, 21, 22]	5	0.8333333333	-0.035166667
7	47	25	17	5.25	6	[5, 8, 10, 15, 21, 25]	5	0.8333333333	-0.035166667
8	49	21	20	4.41	5	[4, 8, 12, 15, 19]	5	1	0.1315
9	51	28	21	5.88	6	[1, 6, 8, 17, 21, 22]	5	0.8333333333	-0.035166667
10	54	33	31	6.93	7	[1, 3, 11, 19, 22, 26, 27]	7	1	0.1315
11	54	33	31	6.93	7	[4, 6, 10, 19, 21, 26, 32]	6	0.8571428571	-0.011357143
12	57	18	17	3.78	4	[9, 13, 15, 16]	4	1	0.1315
13	59	22	22	4.62	5	[4, 5, 6, 9, 12]	5	1	0.1315
14	68	14	14	2.94	3	[1, 2, 8]	3	1	0.1315
15	83	28	22	5.88	6	[2, 6, 7, 15, 17, 23]	5	0.8333333333	-0.035166667
16	83	28	22	5.88	6	[5, 8, 9, 17, 21, 26]	5	0.8333333333	-0.035166667
17	83	28	22	5.88	6	[2, 6, 8, 11, 12, 26]	6	1	0.1315
18	83	28	22	5.88	6	[1, 13, 17, 18, 20, 21]	2	0.3333333333	-0.535166667
19	84	11	9	2.31	3	[2, 8, 9]	1	0.3333333333	-0.535166667
20	90	41	39	8.61	9	[2, 13, 16, 24, 28, 31, 34, 37, 40]	8	0.8888888889	0.0203888889
21	90	41	39	8.61	9	[2, 7, 10, 12, 15, 17, 31, 34, 40]	8	0.8888888889	0.0203888889
22	93	18	18	3.78	4	[2, 8, 11, 18]	4	1	0.1315
23	93	18	18	3.78	4	[1, 6, 10, 17]	4	1	0.1315
24	95	15	13	3.15	4	[5, 7, 9, 13]	3	0.75	-0.1185
25	98	34	34	7.14	8	[5, 6, 7, 9, 16, 22, 25, 30]	8	1	0.1315
26	100	17	13	3.57	4	[6, 7, 13, 15]	4	1	0.1315
27	114	23	16	4.83	5	[5, 8, 15, 17, 23]	3	0.6	-0.2685
28	116	16	16	3.36	4	[7, 11, 12, 13]	4	1	0.1315
29	118	17	17	3.57	4	[4, 5, 15, 16]	4	1	0.1315
30	135	14	10	2.94	3	[3, 4, 11]	3	1	0.1315

Std Dev(y(bar)i-u(hat)pps)
0.1922416906

- $\mu_{(hat)pps}=(1/n)\sum\bar{y}_i$

  $=(1/30)[4/4+4/7+\ldots+3/3]=.8685$
- $V_{(hat)}(\mu_{(hat)pps})=(1/[(n(n-1)])\sum(\bar{y}_i-\mu_{(hat)pps})^2$

  $=(1/[30(29)])[(1-.8685)^2+(.5714-.8685)^2+\ldots(1-.8685)^2]$

  $=(1/30)(.1922^2)0.00123$--->B=0.070
- $\tilde{\iota}_{(hat)pps}=(M/n)\sum\bar{y}_i$

  $=(2209)(.8685)=1918.62$
- $V_{(hat)}(\tilde{\iota}_{(hat)pps})=(M^2/[(n(n-1)])\sum(\bar{y}_i-\mu_{(hat)pps})^2$

  $=(2209^2/[30(29)])[(1-.8685)^2+(.5714-.8685)^2+\ldots(1-.8685)^2]=(2209^2/30)(.1922)^{2=}6008.65$--->B=155.03
- See appendix for SAS code and results

What if a SRS of blocks was used instead of Probabilities Proportional to Size?

	block	plots	single fam
1	11	12	12
2	22	9	9
3	23	10	10
4	31	9	8
5	36	9	8
6	38	18	15
7	43	10	10
8	44	12	11
9	60	15	8
10	62	8	3
11	70	18	16
12	74	17	17
13	79	18	14
14	82	12	11
15	86	16	16
16	88	15	10
17	91	17	17
18	92	18	18
19	95	15	13
20	99	17	14
21	100	17	13
22	106	16	13
23	108	6	4
24	111	10	10
25	113	20	17
26	121	14	14
27	128	14	14
28	129	11	11
29	130	10	10
30	131	7	3

- The macro srs was used to generate a SRS of 30 blocks from the census data. A seed=9101112 was used.
- The macro ratio was used to generate the estimate of the total and proportion of single housing units
- The results were $\tilde{\iota}_{(hat)}$ =1582.13 with B=85.0135 and $\mu_{(hat)}$ =.8725 with B=.0469

Summary of Findings of Total and Proportion of Single Family Units in the South Tabor Neighborhood

	ĩ	µ
census	1919	.8687

	ĩ (Hat)	B (ĩ Hat)	µ (hat)	B (µ hat)	Technical Evaluation
1 Stage PPS	1935.564	93.707	.8762	.042	☺
2 Stage PPS	1918.62	155.03	.8685	.070	☺
1 Stage SRS	1582.13	85.0135	.8725	.0469	☹

Conclusions

The pps method is the favored approach for this dataset. This project strongly supports the statement found in STAT422's text on page 291:

"We now have three estimators of the population total in cluster sampling: the ratio estimator, the unbiased estimator, and the pps estimator. How do we know which is best? Here are some guidelines for answering this question.

If y_i is uncorrelated with m_i, then the unbiased estimator is better than either of the other two.

If y_i is positively correlated with m_i, then the ratio and pps estimators are more precise than the unbiased estimator.

The pps estimator is better than the ratio estimator if the within-cluster variation does not change with changing m_i.

The ratio estimator is better than the pps estimator if the within-cluster variation increases with increasing m_i."

Reference: Elementary Survey Sampling, Scheaffer, Mendenhall III, Ott, 6th Edition, 2006.

Conclusions

- For future studies in obtaining estimates of plot types in neighborhoods where the within cluster variation does not change as the number of plots increases in the blocks, one and two stage clustering with probability proportional to size sampling is recommended
- Two stage clustering would require less measurements as compared with one stage clustering
- When an acceptable method is chosen, the surveyor learned to trust the methods learned in STAT422 within the confidence interval assigned by doing this project