

# Stat 572 Sampling Theory & Practice

## Homework 4

Erik Barry Erhardt

March 29, 2006

### Assignment 4.1.1 *Urban cluster sample.*

In cluster sampling, it is desired that each cluster be representative of the population as a whole.

(a) Use stratified rather than cluster sampling since blocks are self-similar, but different from one another.

(b) Cluster sampling may be reasonable if the constant proportion of nonwhites in each block is near the population proportion of nonwhites. That is, cluster sampling would be ideal if the population of interest is entirely within the sampling frame of blocks.

(c) A situation for cluster sampling, each cluster is as a SRS from the population.

### Assignment 4.1.2 *One-stage cluster sample of journal survey.*

(a) This is a one-stage cluster sample with PSU=scholarly journals in the social and behavioral sciences, and SSU=articles published during 1988 from the selected journals in the PSU.

(b) *Estimate.*

$$\hat{p}_r = \frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n M_i} = \frac{137}{148} = 0.9257$$

$$SE(\hat{p}_r) = \sqrt{\widehat{\text{Var}}(\hat{p}_r)} = \sqrt{\frac{\hat{p}_r(1 - \hat{p}_r)}{n}} = \sqrt{\frac{0.9257(1 - 0.9257)}{26}} = 0.0514$$

Giving a 95% CI for  $p$  of (0.8249, 1) (upper CI limit of 1.0265 truncated at 1).

**(c)** *Ridiculous reasoning.*

Social and behavioral sciences are jumping from a bridge...should our courts of law? The purpose of probability sampling is to get an unbiased picture of the population of interest. A nonprobability sample can make no guarantee about the accuracy of the estimates obtained, since the issue of bias has not been addressed.

Consider a case involving trademark violation or employment discrimination. A nonprobability sample might well include a convenience sample which includes a much higher-than-average proportion of violations or cases of discrimination, or a much lower-than-average proportion (depending whether the prosecution or defense is making the case). Since we SHOULD be interested in making an honest and truthful case, a probability sample done correctly is the only likely way to obtain an accurate estimate of the population. Without truth, we may as well let belief blindly lead.

**Assignment 4.1.3** *Two-stage cluster sample of cans (of worm fragments, yum).*

$$\begin{aligned}\hat{y}_r &= \frac{\sum_{i=1}^n M_i \bar{y}_i}{\sum_{i=1}^n M_i} = 3.6389 \\ \text{SE}(\hat{y}_r) &= \sqrt{\hat{\text{Var}}(\hat{y}_r)} = \sqrt{\frac{1}{n\bar{M}_u^2} \frac{N-n}{N} \frac{\sum_{i=1}^n (M_i \bar{y}_i - M_i \hat{y}_r)^2}{n-1}} \\ &= 0.6102\end{aligned}$$

Giving a 95% CI for  $\bar{y}$  of (2.4429, 4.8348). I would say that's too high, but then the organic corn I eat has worms, but it's no problem off the cob — just avoid those spots.

**Assignment 4.2.1** *Draw sample.*

Table 1 on page 4 gives the list of houses sampled from Lockhart City.

**Assignment 4.2.2** *Estimate and estimate comparison.*

In the table below is the estimate for average price willing to pay, CI and SE(Est) from the current cluster sample, as well as the three previous samples from HW3.

Because the one-stage cluster sampling method samples four consecutive houses, I expect clusters to be self-similar, and clusters to be different from each other. Thus a larger standard error is expected. In the table below, indeed the standard error for cluster sampling is larger than the three other methods.

Sampling Method	Est	95% CI	SE(Est)
One-stage Cluster sample	10.3250	(8.9967,11.6533)	0.6777
Stratified RS with Optimal allocation	9.9653	(9.2597,10.6710)	0.3600
Stratified RS with Proportional allocation	9.3861	(8.6759,10.0964)	0.3624
SRS	9.9250	(8.9761,10.8739)	0.4841

Table 1: 4.2.1 Sample from Lockhart City

Cumul	Dist	House	Cumul	Dist	House	Cumul	Dist	House	Cumul	Dist	House
12478	51	153	18516	59	583	23950	65	579	28056	70	621
12479	51	154	18517	59	584	23951	65	580	28057	70	622
12480	51	155	18534	59	601	23952	65	581	28218	71	24
12481	51	156	18535	59	602	23953	65	582	28219	71	25
12678	51	353	18536	59	603	24070	65	699	28220	71	26
12679	51	354	18537	59	604	24071	65	700	28221	71	27
12680	51	355	18682	59	749	24072	65	701	28502	71	308
12681	51	356	18683	59	750	24073	65	702	28503	71	309
13310	52	460	18684	59	751	24730	66	642	28504	71	310
13311	52	461	18685	59	752	24731	66	643	28505	71	311
13312	52	462	19070	60	219	24732	66	644	28802	71	608
13313	52	463	19071	60	220	24733	66	645	28803	71	609
13846	53	270	19072	60	221	26242	68	617	28804	71	610
13847	53	271	19073	60	222	26243	68	618	28805	71	611
13848	53	272	19222	60	371	26244	68	619	29002	72	86
13849	53	273	19223	60	372	26245	68	620	29003	72	87
14886	55	51	19224	60	373	26470	68	845	29004	72	88
14887	55	52	19225	60	374	26471	68	846	29005	72	89
14888	55	53	19702	61	52	26472	68	847	29122	72	206
14889	55	54	19703	61	53	26473	68	848	29123	72	207
15866	56	478	19704	61	54	26866	69	329	29124	72	208
15867	56	479	19705	61	55	26867	69	330	29125	72	209
15868	56	480	19790	61	140	26868	69	331	29274	72	358
15869	56	481	19791	61	141	26869	69	332	29275	72	359
16462	57	491	19792	61	142	26950	69	413	29276	72	360
16463	57	492	19793	61	143	26951	69	414	29277	72	361
16464	57	493	21634	63	544	26952	69	415	29306	72	390
16465	57	494	21635	63	545	26953	69	416	29307	72	391
16758	57	787	21636	63	546	26994	69	457	29308	72	392
16759	57	788	21637	63	547	26995	69	458	29309	72	393
16760	57	789	22394	63	1304	26996	69	459	29634	72	718
16761	57	790	22395	63	1305	26997	69	460	29635	72	719
17566	58	684	22396	63	1306	27234	69	697	29636	72	720
17567	58	685	22397	63	1307	27235	69	698	29637	72	721
17568	58	686	22542	64	139	27236	69	699	30414	73	745
17569	58	687	22543	64	140	27237	69	700	30415	73	746
17570	58	688	22544	64	141	27342	69	805	30416	73	747
17571	58	689	22545	64	142	27343	69	806	30417	73	748
17572	58	690	22774	64	371	27344	69	807	30782	74	320
17573	58	691	22775	64	372	27345	69	808	30783	74	321
17822	58	940	22776	64	373	27534	70	99	30784	74	322
17823	58	941	22777	64	374	27535	70	100	30785	74	323
17824	58	942	22970	64	567	27536	70	101	30870	74	408
17825	58	943	22971	64	568	27537	70	102	30871	74	409
17862	58	980	22972	64	569	28042	70	607	30872	74	410
17863	58	981	22973	64	570	28043	70	608	30873	74	411
17864	58	982	23414	65	43	28044	70	609	31334	75	147
17865	58	983	23415	65	44	28045	70	610	31335	75	148
18514	59	581	23416	65	45	28054	70	619	31336	75	149
18515	59	582	23417	65	46	28055	70	620	31337	75	150

## Appendix

### code used for the above analysis

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 4.1.2
% NUMEMP = number of empirical studies
% PROB = number of probability samples
% NONPROB = number of non-probability samples
% (not all studies could not be classified as PROB or NONPROB,
% so the sum of the second two columns is not necessarily equal to the first)
%
% NUMEMP,PROB,NONPROB
% one-stage cluster sampling
x=[17,0,17;1,0,1;3,0,3;3,0,2;23,0,19;3,0,3;18,1,16;1,0,1;0,0,0;5,0,4;0,0,0;0,0,0;13,0,13;
  0,0,0;0,0,0;0,0,0;0,0,0;46,0,46;3,0,3;5,0,5;2,0,1;0,0,0;1,0,1;4,2,2;0,0,0;0,0,0];
N=1258;n=26;
Mi=x(:,1);ti=x(:,3);
[sum(ti) sum(Mi)]
p_r=sum(ti)/sum(Mi)
var_p_r=p_r*(1-p_r)/n; se_p=sqrt(var_p_r)
alpha=0.05; CI_p=[p_r-norminv(1-alpha/2)*se_p, p_r+norminv(1-alpha/2)*se_p]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 4.1.3
% two-stage cluster sampling
x=[1 4 0 3 4 0 5 3 7 3 4 0;5 2 1 6 9 7 5 0 3 1 7 0;7 4 2 6 8 3 1 2 5 4 9 0]';
N=580;n=12;Mi=24*ones(n,1);mi=3*ones(n,1);
ybar_i=mean(x,2);
ybar_r=sum(Mi.*ybar_i)./sum(Mi)
si=sqrt(var(x'))';
term1_clusters=((N-n)/N)*sum((Mi.*ybar_i-Mi*ybar_r).^2)/((n-1)*n);
term2_within_clusters=(1/(n*N))*sum(Mi.^2.*((Mi-mi)./Mi).*(si.^2./mi));
var_ybar_r=(1/mean(Mi)^2) * ( term1_clusters + term2_within_clusters );
se_ybar_r=sqrt(var_ybar_r)
alpha=0.05; CI_ybar=[ybar_r-norminv(1-alpha/2)*se_ybar_r, ybar_r+norminv(1-alpha/2)*se_ybar_r]
% SRS comparison
x=[1 4 0 3 4 0 5 3 7 3 4 0 5 2 1 6 9 7 5 0 3 1 7 0 7 4 2 6 8 3 1 2 5 4 9 0]';
yb=mean(x)
se_yb=sqrt(var(x)/length(x))
[yb-1.96*se_yb, yb+1.96*se_yb ]
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 4.2.1
rand('seed',2718281828459045)
n=50;K=19664;M=4;
x=randperm(K/M)';
x=x(1:n);
x=sort(x)*4;
y=zeros(n*M,1);
for i=1:n;
    y((i-1)*M+1):(i*M)=[(x(i)-3):x(i)];
end;
y=y+12325;
cumulative_house_numbers=y;
[district house]=cumulative_to_district_house(cumulative_house_numbers);
cross_district=0;
for i=1:n; if district(4*i-3,1) ~= district(4*i,1); cross_district=cross_district+1; end; end;
cross_district % if 0, then no clusters cross district lines.
dh=[district house]';
fid=fopen('hw4_ad21','w');
fprintf(fid,'%d %d\n',dh);
fprintf(fid,'0 0 99999');
fclose(fid);
type hw4_ad21
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 4.2.2
./survey
hw4_ad21
hw4_sa21
0 0 0
% cut off first and last lines, then...

```

```

x1=load('hw4_sa21');
price=x1(:,7);
price = reshape(price,M,n)';
price_mean=mean(price,2);
price_var =var(price)';
% one-stage cluster version
% to use the code from 4.1.3
N=K/M;
ybar_i=price_mean;
ybar_r=sum(M*ybar_i)/(n*M)
var_ybar_r=((N-n)/N) * var(price_mean)/(n);
se_ybar_r=sqrt(var_ybar_r)
alpha=0.05; CI_ybar=[ybar_r-norminv(1-alpha/2)*se_ybar_r, ybar_r+norminv(1-alpha/2)*se_ybar_r]

% two-stage cluster version
% to use the code from 4.1.3
Mi=M*ones(n,1);mi=M*ones(n,1);
ybar_i=price_mean;
ybar_r=sum(Mi.*ybar_i)./sum(Mi)
si=sqrt(price_var);
term1_clusters=((N-n)/N)*sum((Mi.*ybar_i-Mi*ybar_r).^2)/((n-1)*n);
term2_within_clusters=(1/(n*N))*sum(Mi.^2.*((Mi-mi)./Mi).*(si.^2./mi));
var_ybar_r=(1/mean(Mi)^2) * ( term1_clusters + term2_within_clusters );
se_ybar_r=sqrt(var_ybar_r)
alpha=0.05; CI_ybar=[ybar_r-norminv(1-alpha/2)*se_ybar_r, ybar_r+norminv(1-alpha/2)*se_ybar_r]

function [district, house] = cumulative_to_district_house(cumulative_house_numbers)
% For use with Sharon L. Lohr's "Sampling: Design and Analysis"
%
% This function is useful for translating cumulative house numbers
% from Stephens County to their district and house number.
% Run this on a list of cumulative house numbers,
% then feed the district and house numbers into the "survey" program.
% The output from this program mimics that of the "addgen" program,
% except that you have generated the sample in your
% cumulative_house_numbers list.
%
% Written: 3/25/2006
%
% Erik Barry Erhardt, M.S., Ph.D. Student Statistics
% Dept of Math & Stats, Univ. of New Mexico, Albuquerque, NM 87131, U.S.A.
% Office: Humanities 328, MSC03 2150
% erike AT stat.unm.edu http://www.stat.unm.edu/~erike/
% Office: (505)277-0757 Fax: (505)277-5505
% error checking
[r,c]=size(cumulative_house_numbers);
if r ~= 1 & c ~= 1
    error('Input must be a vector');
    return;
end;
if c>r; cumulative_house_numbers=cumulative_house_numbers'; end; % make a column vector
n=length(cumulative_house_numbers); % how many houses
% From Figure A.2 on p. 416 of Lohr
% Stephens County district information
% columns are:
% districts,
% number of houses,
% cumulative house count,
% population,
% mean assessed house valuation
Stephens_County=[
1 142 142 526 65248. ;...
2 153 295 624 58759. ;...
3 135 430 508 62319. ;...
4 128 558 560 59416. ;...
5 110 668 455 57202. ;...
6 103 771 404 59290. ;...
7 105 876 421 71122. ;...
8 385 1261 1488 79265. ;...
9 296 1557 1112 75921. ;...
10 287 1844 994 68254. ;...
11 253 2097 929 60660. ;...
12 172 2269 628 53569. ;...
13 198 2467 768 65182. ;...
14 432 2899 1595 77907. ;...

```

```

15 248 3147 864 65739. ;...
16 251 3398 915 53771. ;...
17 221 3619 864 68257. ;...
18 297 3916 1099 78449. ;...
19 235 4151 812 70772. ;...
20 171 4322 687 52711. ;...
21 135 4457 525 66739. ;...
22 254 4711 923 66249. ;...
23 203 4914 708 74757. ;...
24 244 5158 825 75766. ;...
25 202 5360 799 68989. ;...
26 103 5463 388 56994. ;...
27 102 5565 398 58940. ;...
28 115 5680 448 60448. ;...
29 180 5860 693 69111. ;...
30 190 6050 766 69685. ;...
31 152 6202 633 70276. ;...
32 141 6343 572 63819. ;...
33 143 6486 610 58636. ;...
34 135 6621 491 55554. ;...
35 178 6799 699 62361. ;...
36 221 7020 811 60052. ;...
37 174 7194 719 55699. ;...
38 101 7295 390 53322. ;...
39 95 7390 312 57174. ;...
40 130 7520 446 55702. ;...
41 152 7672 533 53285. ;...
42 169 7841 672 56866. ;...
43 91 7932 371 50710. ;...
44 283 8215 1029 60057. ;...
45 562 8777 2079 57233. ;...
46 312 9089 1149 52719. ;...
47 897 9986 3263 62034. ;...
48 734 10720 2623 60764. ;...
49 963 11683 3490 60010. ;...
50 642 12325 2318 54498. ;...
51 525 12850 1825 95123. ;...
52 726 13576 2497 68406. ;...
53 674 14250 1948 53634. ;...
54 585 14835 1219 48643. ;...
55 553 15388 1090 43493. ;...
56 583 15971 1977 95110. ;...
57 911 16882 2691 84394. ;...
58 1051 17933 2663 57657. ;...
59 918 18851 1824 36706. ;...
60 799 19650 1636 44308. ;...
61 545 20195 1853 101906. ;...
62 895 21090 2588 74815. ;...
63 1313 22403 2642 55560. ;...
64 968 23371 2457 62813. ;...
65 717 24088 2203 69846. ;...
66 651 24739 2197 93771. ;...
67 886 25625 2711 82902. ;...
68 912 26537 2750 76832. ;...
69 898 27435 2671 72062. ;...
70 759 28194 2650 79887. ;...
71 722 28916 2568 87383. ;...
72 753 29669 2652 80341. ;...
73 793 30462 2763 79833. ;...
74 725 31187 2560 83354. ;...
75 802 31989 2870 80522. ];

SC_district = Stephens_County(:,1); % give specific names to the relevant
SC_houses = Stephens_County(:,2); % columns of the Stephens_County matrix
SC_cum_houses = Stephens_County(:,3); % of

district = zeros(n,1); house = zeros(n,1); % prespecify the district and house vectors
% assign the district and house numbers
for i=1:n;
    % Find district number by finding
    % the minimum row for which the cumulative_house_numbers
    % is less than the Stephens County cumulative house number
    % and setting the associated Stephens County district to the district.
    SC_row = min(find((cumulative_house_numbers(i) < (SC_cum_houses+1))));
    district(i) = SC_district(SC_row);
    % start_house is the cumulative house number from the previous district,
    % or the 0th house of the current district.
    % To avoid problems with the first district, I don't use the previous district's cumulative house number.
    start_house = SC_cum_houses(SC_row)-SC_houses(SC_row);
    % making the current house our cumulative_house_numbers minus the starting house.
    house(i) = cumulative_house_numbers(i)-start_house;
end;

% EOF

```