

Stat 572 Sampling Theory & Practice

Homework 3 Part 2

Erik Erhardt and Nina Greenberg

February 15, 2006

All relevant code is included in the appendix.

Assignment 3.2.1 *Stratification.*

Selected strata were based on house value because the price a household is willing to pay for cable may be proportional to the value of their house. We note such stratification is unlikely to group households by presence or absence of children under 11. However, as the total number of children per household is expected to decrease with increased socio-economic factors, stratification based on house evaluation may be useful in determining the related question of total numbers of children per household.

Particular strata were identified by observing patterns in mean house evaluations for each district and grouping mostly contiguous areas. Different valuation groupings were selected for Lockhart City and Stephens County. In all, eight strata were selected (Table 1) with 5 (No.1–5) in Lockhart City and 3 (No. 6–8) in surrounding rural areas of Stephens County. The plot in Figure 1 on page 2 illustrates the selected stratification.

Strata	House Value (\$1,000)	Number of Households	Porportion of Lockhart City
1	35–55	3529	.1795
2	55–70	4775	.2428
3	70–80	4257	.2165
4	80–85	4077	.2073
5	85–105	3026	.1539
6	50–60	3999	
7	60–70	5799	
8	70–80	2527	

Table 1: Strata criteria.

FIGURE A.1
A district map of Stephens County

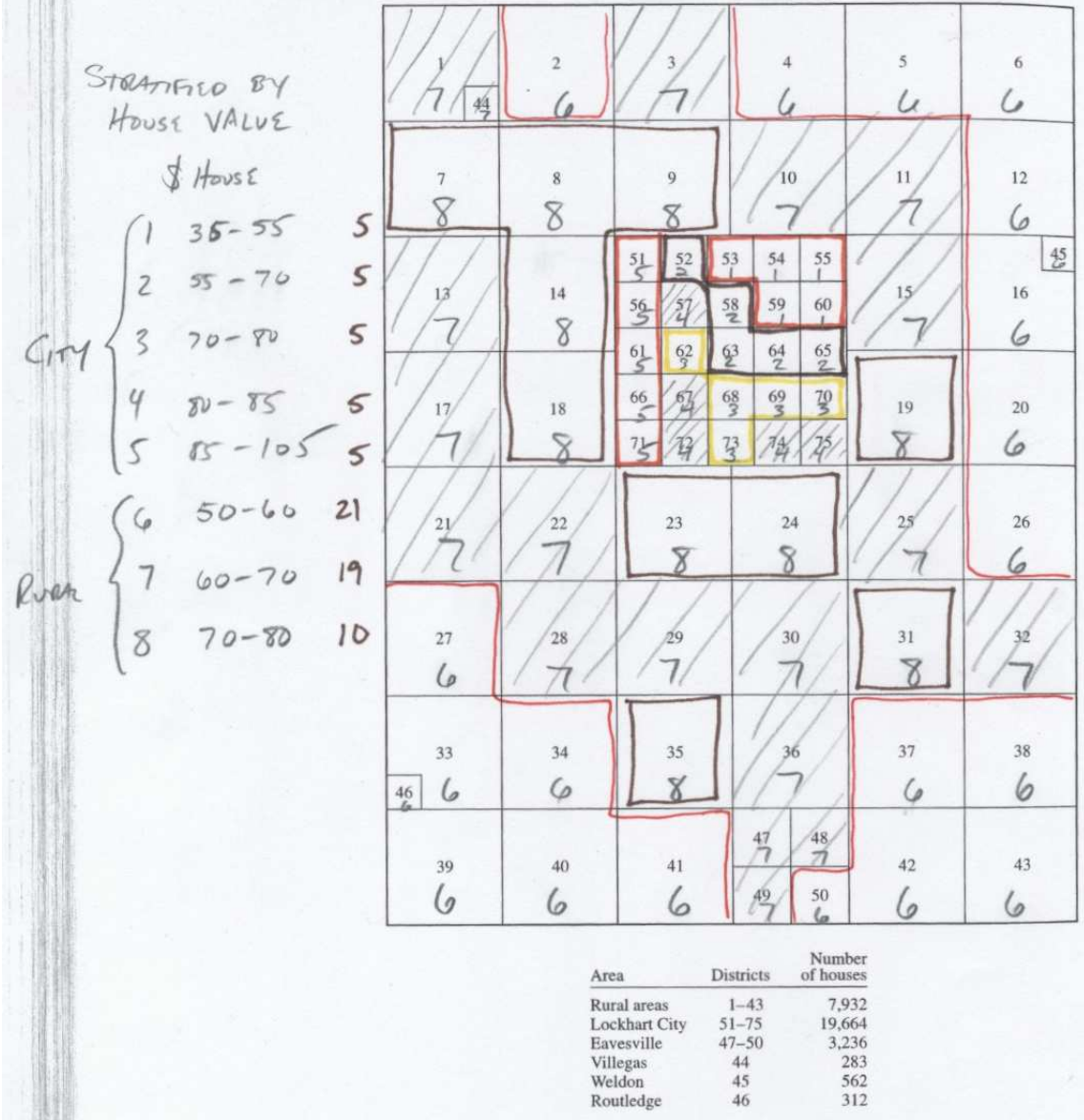


Figure 1: 3.2.1. Stratification.

Assignment 3.2.2 *Stratified RS with Proportional allocation.*

Based on strata population sizes $N = (3529, 4775, 4257, 4077, 3026)$, we chose proportional sample sizes of $n = (36, 49, 43, 41, 31)$. The table below gives the estimates (sample mean and proportion) with a 95% CI, and the standard error of the estimate.

	Est	95% CI	SE(Est)
Price	9.3861	(8.6759, 10.0964)	0.3624
≤ 11 yo.	0.3755	(0.3086, 0.4424)	0.0341

Assignment 3.2.3 *SRS*

For a SRS of $n = 200$, the table below gives the estimates (sample mean and proportion) with a 95% CI, and the standard error of the estimate.

The standard error for the mean is much larger (about 33%), which is expected because that response is related to our stratification variable. The proportion standard error is not different because stratification had no effect on that variable.

	Est	95% CI	SE(Est)
Price	9.9250	(8.9761, 10.8739)	0.4841
≤ 11 yo.	0.3600	(0.2936, 0.4264)	0.0339

Assignment 3.2.4 *Stratified RS with Optimal allocation.*

Based on strata population sizes and sample variances for each strata, we calculate optimal sample sizes and round for $n_{\text{opt}} = (28, 51, 46, 38, 37)$. We note the sample sizes for the middle strata are similar to the proportional allocation, 8 less in the first strata, and 6 more in the last strata. The table below gives the estimates (sample mean and proportion) with a 95% CI, and the standard error of the estimate.

For Price, the SE for the optimal allocation differs by 0.0024, which is not effectively different than our proportional allocation. The 95% is wider by only a penny for price. For proportion ≤ 11 yo., the SE differs by only 0.0004, a negligible difference.

	Est	95% CI	SE(Est)
Price	9.9653	(9.2597, 10.6710)	0.3600
≤ 11 yo.	0.3901	(0.3245, 0.4558)	0.0335

Assignment 3.2.5 *Optimal vs. proportional allocation.*

Optimal allocation is expected to perform better than proportional allocation when the variance within strata differ.

Based on our stratification, this did not occur in Lockhart City. Optimal allocation did not give different standard errors from proportional allocation because each of our strata have similar variances (see 3.2.6).

Assignment 3.2.6 *Deficiencies in stratification.*

There does not seem to be any deficiencies in our stratification for the purposes of reducing variance in the estimate of price willing to pay for cable. Levene's test for unequal variances shows that our strata variances are not different from each other (see "Price willing to pay for cable" **results below**). That indicates that we chose strata that are self similar in the price willing to pay for cable. The ANOVA below strongly indicates differences between strata (p-value < 0.001) (see the **results below**).

There are deficiencies in our stratification for the purposes of reducing variance in the estimate of proportion of children aged 11 and younger. This is because our stratification variable of average house value is not related to whether the household has children. This could be improved by also stratifying on the average number of people per household, with larger numbers being more likely to have children aged 11 and younger. Levene's test for unequal variances shows that our strata variances are not different from each other (we understand that this test does not apply since the data do not come from a continuous distribution) (see "Proportion of households with children 11 years and younger" **results below**). That indicates that we chose strata that are self similar in the proportion of households with children aged 11 and younger. The ANOVA below indicates no difference between strata (p-value = 0.333) (see the **results below**).

Negligible differences in SE in optimal allocation versus proportional allocation is attributed to our strata having similar variances. However, if we observed a stratum that had a much larger variance than the other strata, then observations in that strata are less self-similar than in other strata, indicating that the stratification scheme may need revision.

Price willing to pay for cable

Test for Equal Variances: price versus strata

95% Bonferroni confidence intervals for standard deviations

strata	N	Lower	StDev	Upper
1	36	3.07180	4.03113	5.75175
2	49	4.24150	5.37101	7.22715
3	43	4.20229	5.39934	7.43689
4	41	3.58083	4.62628	6.42995
5	31	4.62238	6.18270	9.12029

Bartlett's Test (normal distribution)

Test statistic = 7.01, p-value = 0.135

Levene's Test (any continuous distribution)

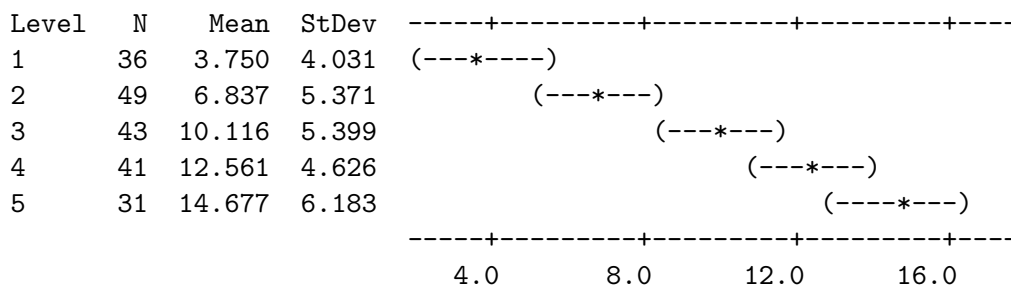
Test statistic = 0.61, p-value = 0.659

One-way ANOVA: price versus strata

Source	DF	SS	MS	F	P
strata	4	2766.1	691.5	26.03	0.000
Error	195	5180.7	26.6		
Total	199	7946.9			

S = 5.154 R-Sq = 34.81% R-Sq(adj) = 33.47%

Individual 95% CIs For Mean Based on Pooled StDev



Pooled StDev = 5.154

Proportion of households with children 11 years and younger.

Test for Equal Variances: child versus strata

95% Bonferroni confidence intervals for standard deviations

strata	N	Lower	StDev	Upper
1	36	0.364315	0.478091	0.682157
2	49	0.360448	0.456435	0.614172
3	43	0.393647	0.505781	0.696647
4	41	0.386065	0.498779	0.693241
5	31	0.363628	0.486373	0.717464

Bartlett's Test (normal distribution)

Test statistic = 0.56, p-value = 0.967

Levene's Test (any continuous distribution)

Test statistic = 1.15, p-value = 0.333

One-way ANOVA: child versus strata

Source	DF	SS	MS	F	P
strata	4	1.083	0.271	1.15	0.333
Error	195	45.792	0.235		
Total	199	46.875			

S = 0.4846 R-Sq = 2.31% R-Sq(adj) = 0.31%

Individual 95% CIs For Mean Based on Pooled StDev

Level	N	Mean	StDev	
1	36	0.3333	0.4781	+-----+-----+-----+----- (-----*-----)
2	49	0.2857	0.4564	(-----*-----)
3	43	0.4884	0.5058	(-----*-----)
4	41	0.4146	0.4988	(-----*-----)
5	31	0.3548	0.4864	(-----*-----)

0.15 0.30 0.45 0.60

Pooled StDev = 0.4846

Appendix

code used for the above analysis

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% 3.2.2
./addgen
hw3_ad11
908318
53-55,59,60
36
no
./addgen
hw3_ad12
146687
52,58,63-65
49
no
./addgen
hw3_ad13
514701
62,68-70,73
43
no
./addgen
hw3_ad14
405809
57,67,72,74-75
41
no
./addgen
hw3_ad15
733812
51,56,61,66,71
31
no

./survey
hw3_ad11
hw3_sa11
0 0 0
./survey
hw3_ad12
hw3_sa12
0 0 0
./survey
hw3_ad13
hw3_sa13
0 0 0
./survey
hw3_ad14
hw3_sa14
0 0 0
./survey
hw3_ad15
hw3_sa15
0 0 0

% edit hw3_sa1x removing first and last line, and saving hw3_sa1x.txt
% matlab
x1=load('hw3_sa11.txt');
x2=load('hw3_sa12.txt');
x3=load('hw3_sa13.txt');
x4=load('hw3_sa14.txt');
x5=load('hw3_sa15.txt');

x1=[x1 1*ones(length(x1),1)]; % indicator variables
x2=[x2 2*ones(length(x2),1)];
x3=[x3 3*ones(length(x3),1)];
x4=[x4 4*ones(length(x4),1)];
x5=[x5 5*ones(length(x5),1)];

x=[x1;x2;x3;x4;x5];
y=[x(:,7), ones(length(x),1).*(x(:,5)>0), x(:,end)]; % convert to binary variable
N=[3529 4775 4257 4077 3026];
n=[length(find(y(:,3)==1))...
length(find(y(:,3)==2))...
length(find(y(:,3)==3))...
length(find(y(:,3)==4))...
length(find(y(:,3)==5))]
mu=[mean(y(find(y(:,3)==1),1))...
mean(y(find(y(:,3)==2),1))...
mean(y(find(y(:,3)==3),1))...
mean(y(find(y(:,3)==4),1))...
mean(y(find(y(:,3)==5),1))]

```

```

mus=[std(y(find(y(:,3)==1),1))...
      std(y(find(y(:,3)==2),1))...
      std(y(find(y(:,3)==3),1))...
      std(y(find(y(:,3)==4),1))...
      std(y(find(y(:,3)==5),1))]

p =[mean(y(find(y(:,3)==1),2))...
    mean(y(find(y(:,3)==2),2))...
    mean(y(find(y(:,3)==3),2))...
    mean(y(find(y(:,3)==4),2))...
    mean(y(find(y(:,3)==5),2))]

mustr=sum((N/sum(N)).*mu)
pstr=sum((N/sum(N)).*p)
correction_term=(1-n./N).*(N/sum(N)).^2;
mustr_se=sqrt(sum(correction_term.*(mus.^2)./n))
pstr_se=sqrt(sum(correction_term.*p.*(1-p)./(n-1)))
alpha=0.05;
z=norminv(1-alpha/2);
ci=[mustr mustr-z*mustr_se mustr+z*mustr_se;...
    pstr pstr-z*pstr_se pstr+z*pstr_se]
[mustr_se;pstr_se]

% n =
%    36    49    43    41    31
% mu =
%    3.7500    6.8367    10.1163    12.5610    14.6774
% mus =
%    4.0311    5.3710    5.3993    4.6263    6.1827
% p =
%    0.3333    0.2857    0.4884    0.4146    0.3548
% mustr =
%    9.3861
% pstr =
%    0.3755
% mustr_se =
%    0.3624
% pstr_se =
%    0.0341
% ci =
%    9.3861    8.6759    10.0964
%    0.3755    0.3086    0.4424
% ans =
%    0.3624
%    0.0341

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3.2.3
./addgen
hw3_ad3
339362
51-75
200
no
./survey
hw3_ad3
hw3_sa3
0 0 0

% edit hw3_sa1x removing first and last line, and saving hw3_sa1x.txt
% matlab
x=load('hw3_sa3.txt');
y=[x(:,7), ones(length(x),1).*(x(:,5)>0)]; % convert to binary variable
N=19664;
n=length(y);
mu=mean(y(:,1))
mus=std(y(:,1))
p=mean(y(:,2))
mustr=mu
pstr=p

correction_term=(N-n)/N; % finite population correction
mustr_se=sqrt(sum(correction_term.*(mus.^2)./n))
pstr_se=sqrt(sum(correction_term.*p.*(1-p)./(n-1)))
alpha=0.05;
z=norminv(1-alpha/2);
ci=[mustr mustr-z*mustr_se mustr+z*mustr_se;...
    pstr pstr-z*pstr_se pstr+z*pstr_se]
[mustr_se;pstr_se]

```

```

% mu =
% 9.9250
% mus =
% 6.8816
% p =
% 0.3600
% mustr =
% 9.9250
% pstr =
% 0.3600
% mustr_se =
% 0.4841
% pstr_se =
% 0.0339
% ci =
% 9.9250      8.9761      10.8739
% 0.3600      0.2936      0.4264
% ans =
% 0.4841
% 0.0339

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% 3.2.4

% (a)
%% using mus and N from 2.2
% using Neyman allocation (assumes cost per observation constant)
n_l=round(sum(n)*N.*mus/sum(N.*mus))
sum(n_l)
% n_l =
% 28    51    46    38    37

% (b)
./addgen
hw3_ad41
995466
53-55,59,60
28
no
./addgen
hw3_ad42
200340
52,58,63-65
51
no
./addgen
hw3_ad43
798070
62,68-70,73
46
no
./addgen
hw3_ad44
951898
57,67,72,74-75
38
no
./addgen
hw3_ad45
220978
51,56,61,66,71
37
no

./survey
hw3_ad41
hw3_sa41
0 0 0
./survey
hw3_ad42
hw3_sa42
0 0 0
./survey
hw3_ad43
hw3_sa43
0 0 0
./survey
hw3_ad44
hw3_sa44
0 0 0
./survey
hw3_ad45
hw3_sa45
0 0 0

% edit hw3_sa1x removing first and last line, and saving hw3_sa1x.txt

```

```

% matlab
x1=load('hw3_sa41.txt');
x2=load('hw3_sa42.txt');
x3=load('hw3_sa43.txt');
x4=load('hw3_sa44.txt');
x5=load('hw3_sa45.txt');

x1=[x1 1*ones(length(x1),1)]; % indicator variables
x2=[x2 2*ones(length(x2),1)];
x3=[x3 3*ones(length(x3),1)];
x4=[x4 4*ones(length(x4),1)];
x5=[x5 5*ones(length(x5),1)];

x=[x1;x2;x3;x4;x5];

y=[x(:,7), ones(length(x),1).*(x(:,5)>0), x(:,end)]; % convert to binary variable
N=[3529 4775 4257 4077 3026];
n=[length(find(y(:,3)==1))...
    length(find(y(:,3)==2))...
    length(find(y(:,3)==3))...
    length(find(y(:,3)==4))...
    length(find(y(:,3)==5))]
mu=[mean(y(find(y(:,3)==1),1))...
    mean(y(find(y(:,3)==2),1))...
    mean(y(find(y(:,3)==3),1))...
    mean(y(find(y(:,3)==4),1))...
    mean(y(find(y(:,3)==5),1))]
mus=[std(y(find(y(:,3)==1),1))...
    std(y(find(y(:,3)==2),1))...
    std(y(find(y(:,3)==3),1))...
    std(y(find(y(:,3)==4),1))...
    std(y(find(y(:,3)==5),1))]
p = [mean(y(find(y(:,3)==1),2))...
    mean(y(find(y(:,3)==2),2))...
    mean(y(find(y(:,3)==3),2))...
    mean(y(find(y(:,3)==4),2))...
    mean(y(find(y(:,3)==5),2))]

mustr=sum((N/sum(N)).*mu)
pstr=sum((N/sum(N)).*p)

correction_term=(1-n./N).*(N/sum(N)).^2;
mustr_se=sqrt(sum(correction_term .* (mus.^2)./n))
pstr_se =sqrt(sum(correction_term .* p.*(1-p)./(n-1)))

alpha=0.05;
z=norminv(1-alpha/2);

ci=[mustr mustr-z*mustr_se mustr+z*mustr_se;...
    pstr pstr-z*pstr_se pstr+z*pstr_se]
[mustr_se;pstr_se]

% n =
% 28 51 46 38 37
% mu =
% 2.3214 8.2353 10.5435 12.7632 17.0270
% mus =
% 3.1862 5.4611 4.8566 6.0065 5.5851
% p =
% 0.1429 0.4706 0.3696 0.5000 0.4324
% mustr =
% 9.9653
% pstr =
% 0.3901
% mustr_se =
% 0.3600
% pstr_se =
% 0.0335
% ci =
% 9.9653 9.2597 10.6710
% 0.3901 0.3245 0.4558
% ans =
% 0.3600
% 0.0335

```