

Stat 572 Sampling Theory & Practice

Homework 3 Part 1&3

Erik Barry Erhardt

February 12, 2006

Assignment 3.1.1 *Estimator \hat{p}_{str} .*

$$\hat{p}_{str} = \sum_{l=1}^L \left(\frac{N_l}{N} \right) \hat{p}_l,$$

where N is population size, N_l is stratum l size, and \hat{p}_l is the stratum l proportion.

$$E(\hat{p}_{str}) = E \left(\sum_{l=1}^L \left(\frac{N_l}{N} \right) \hat{p}_l \right) = \sum_{l=1}^L \left(\frac{N_l}{N} \right) E(\hat{p}_l) = \sum_{l=1}^L \left(\frac{N_l}{N} \right) p_l = p$$

$$\text{Var}(\hat{p}_{str}) = \text{Var} \left(\sum_{l=1}^L \left(\frac{N_l}{N} \right) \hat{p}_l \right) = \sum_{l=1}^L \left(\frac{N_l}{N} \right)^2 \text{Var}(\hat{p}_l)$$

$$\widehat{\text{Var}}(\hat{p}_{str}) = \sum_{l=1}^L \left(\frac{N_l}{N} \right)^2 \frac{\hat{p}_l(1 - \hat{p}_l)}{n_l - 1}$$

Note that \hat{p}_{str} is unbiased for p since $E(\hat{p}_{str}) = p$.

Assignment 3.1.2 *Stratification variables.*

(a) *NYC HW computer programming classes.*

Five boroughs of NYC.

(b) *NM Rail Runner commuter train.*

Living location: By distance from ABQ, and distance from closest station.

Assignment 3.1.3 *Total number of farm acres planted in trees for a state.*

Table 1 on page 2 give the sample means and variances for each stratum.

The estimate of the total number of acres is $\hat{t}_{\text{str}} = \sum_{l=1}^L N_l \bar{y}_l$, with variance $\widehat{\text{Var}}(\hat{t}_{\text{str}}) = \sum_{l=1}^L N_l^2 \frac{N_l - n_l}{N - l} \frac{s_l^2}{n_l}$. Confidence interval for the population total is $\hat{t}_{\text{str}} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\hat{t}_{\text{str}})}$.

$\hat{t}_{\text{str}} = 50441.68$ with 95% CI: (41930.50, 58952.87).

	Stratum I	Stratum II	Stratum III	Stratum IV
N_l	86	72	52	30
n_l	14	12	9	5
\bar{y}_l	63.29	183	339.44	472.4
$\text{Var}(y_{il})$	1077.45	9054.18	17166.78	72376.3

Table 1: 3.1.3

The plot in Figure 1 on page 2 provides a boxplot of the samples showing both an increasing mean and variation from Strata I to Strata IV.

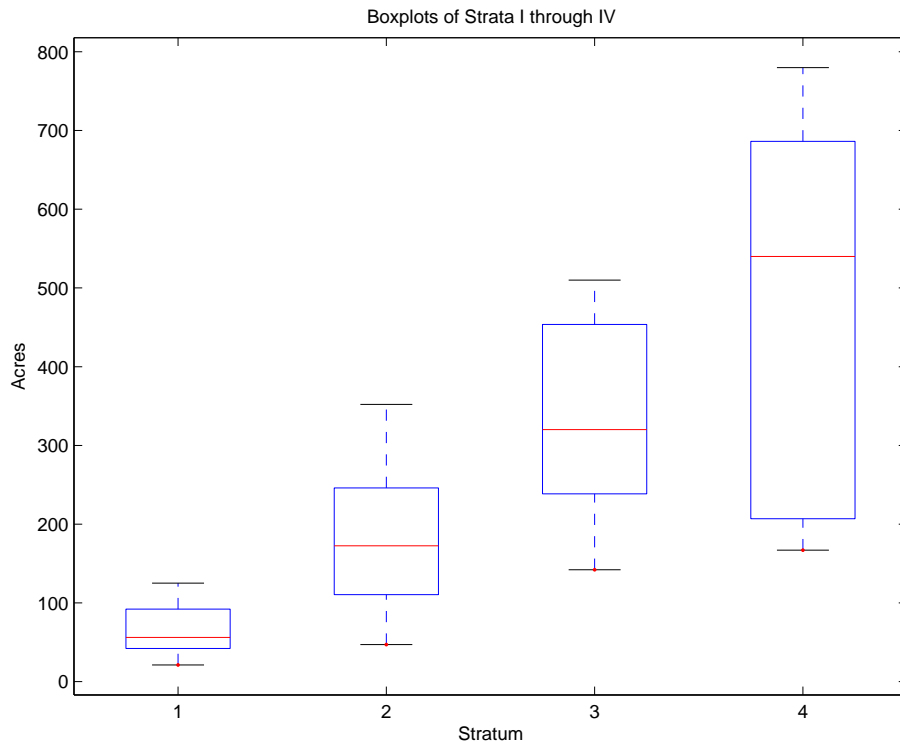


Figure 1: Boxplot of samples showing increasing mean and variation.

Assignment 3.1.4 *Neyman allocation.*

For estimation error bound of $B = 5000$, we first calculate total sample size with

$$n = \frac{(\sum_{l=1}^L N_l s_l)^2}{\frac{B^2}{z_{\frac{\alpha}{2}}} + \sum_{l=1}^L N_l s_l^2},$$

then allocation the individual strata sample sizes with

$$n_l = n \frac{N_l s_l}{\sum_{k=1}^L N_k s_k}.$$

Solving gives $n = 58.46$, which we will round up to $n = 59$ after allocation. Solving for strata gives $n_1 = 6.72, n_2 = 16.31, n_3 = 16.22, n_4 = 19.21$ rounding gives $n_1 = 7, n_2 = 17, n_3 = 16, n_4 = 19$ totaling $n = 59$.

Assignment 3.3.1 *Extra credit.*

Appendix

code used for the above analysis

```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3.1.3
e=zeros(4,4);
x1=[ 97  67  42 125  25  92 105  86  26  43  45  59  53  21]; e(1,:)= [86,length(x1),mean(x1),var(x1)];
x2=[125 155  67  96 256  47 310 236 220 352 142 190]; e(2,:)= [72,length(x2),mean(x2),var(x2)];
x3=[142 256 310 440 495 510 320 396 186]; e(3,:)= [52,length(x3),mean(x3),var(x3)];
x4=[167 655 220 540 780]; e(4,:)= [30,length(x4),mean(x4),var(x4)];
e'
t_str=sum(e(:,1).*e(:,3))
Var_t_str=sum(e(:,1).^2.*((e(:,1)-e(:,2))./e(:,1)).*e(:,4)./e(:,2))
alpha=0.05;z=norminv(1-alpha/2);
est=t_str;se=sqrt(Var_t_str);
lci=est-z*se;
uci=est+z*se;
[est, lci, uci]

x1=[ 97  67  42 125  25  92 105  86  26  43  45  59  53  21];
x2=[125 155  67  96 256  47 310 236 220 352 142 190 NaN NaN];
x3=[142 256 310 440 495 510 320 396 186 NaN NaN NaN NaN];
x4=[167 655 220 540 780 NaN NaN NaN NaN NaN NaN NaN NaN];

x=[x1;x2;x3;x4]';
boxplot(x);title('Boxplots of Strata I through IV');xlabel('Stratum');ylabel('Acres');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3.1.4
B=5000;
num=(sum(e(:,1).*sqrt(e(:,4))))^2;
den1=B^2/z^2;
den2=sum(e(:,1).*e(:,4));
n=num/(den1+den2)
n_l=n*(e(:,1).*sqrt(e(:,4)))/sum(e(:,1).*sqrt(e(:,4)))

```