36-303 Sampling, Surveys & Society
Homework 05 Solutions

April 11, 2012

1 True/False

(a) True: we aim for homogeneity within strata, so that we don't introduce additional error by taking a sample from within each stratum.
(b) False: we aim for heterogeneity within clusters, so that we don't introduce additional error by taking a sample of the clusters.

2 Stratifying Variables

Answers may vary; an acceptable set follows.
(a) gender, income level, reported race/ethnicity, reported party affiliation
(b) class year, major
(c) specific time slot, type of program (sitcom, drama, news, etc.)

3 Stratified Sampling

(a) In order to estimate the total number of bushels of clams in the area, we will estimate the average number in a tow's worth of area, $\bar{y}_{st}$ and then multiply afterwards by the number of tows that would be needed to cover the entire area. We calculate some intermediate quantities from the data: N_h, N, W_h, and f_h, as defined on page 2 of the stratification handout.

Below is a table with the relevant quantities for the calculation. Note that when I did my calculation, I didn't round until the end, but some of the quantities below might be reported rounded off for space reasons.

Stratum	Area	N_h	W_h	n_h	$\bar{y}_h$	f_h	s_h^2
1	222.81	5704	0.428	4	0.44	0.0007	0.068
2	49.61	1270	0.095	6	1.17	0.0047	0.042
3	50.25	1286	0.097	3	3.92	0.0023	2.146
4	197.81	5064	0.380	5	1.80	0.0010	0.794
Total	520.48	13324	1	18	?	0.0014	?

To estimate the average number of bushels in a tow's worth of area, using the information in the table, we have

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h = 1.36 \tag{1}$$

To get the total number of bushels in the area, we multiply this estimate by N $\approx$ 13324, yielding $\approx$ 18153.

For the standard error of our estimate, we first calculate the variance of the average number of bushels in a tow's worth of area.

$$Var(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{s_h^2}{n_h} = 0.033 \tag{2}$$

Note that for the total number of bushels,

$$\sqrt{Var(N\bar{y}_{st})} = \sqrt{N^2 Var(\bar{y}_{st})} = N\sqrt{Var(\bar{y}_{st})} \approx 2411 \tag{3}$$

(b) We are given that $s^2 = 1.084$.

$$Var(\bar{y}_{SRS}) = (1 - f)\frac{s^2}{n} = (1 - \frac{18}{13324})\frac{1.084}{18} \approx 0.060 \tag{4}$$

.

Thus the design effect, $\frac{Var(\bar{y}_{st})}{Var(\bar{y}_{SRS})}$, is ≈ 0.544, so it was indeed better to do a stratified sample; it decreased our standard error of estimation by about 46%.

4 Clustered Sampling

(a) This is a clustered sample because an SRS of 3-packs (clusters, psu's) was taken, instead of a sample of cans (units, ssu's).
(b) Each cluster's mean is given below.

Cluster	Mean
1	4.33
2	3.33
3	1
4	5
5	7
6	3.33
7	3.67
8	1.67
9	5
10	2.67
11	6.67
12	0
Overall	3.64

For the clustered sample, we have

$$\bar{y}_{cl} = \frac{1}{12}\sum_{i=1}^{12}\bar{y}_i = 3.64 \tag{5}$$

and

$$s^2_{\bar{y}_i} = \frac{1}{12-1}\sum_{i=1}^{12}(\bar{y}_i - \bar{y}_{cl})^2 \tag{6}$$

$$= \frac{1}{11}[(4.33-3.64)^2 + (3.33-3.64)^2 + ... + (0-3.64)^2] \approx \ 4.53 \tag{7}$$

Hence,

$$Var(\bar{y}_{cl}) = (1-\frac{n}{N})\frac{1}{n}s^2_{\bar{y}_i} = (1-\frac{12}{580})\frac{1}{12}4.53 \approx 0.37 \tag{8}$$

So the standard error of our estimate, the square root of $Var(\bar{y}_{cl})$, is ≈ 0.61.
(c) We have that the design effect

$$\frac{Var(\bar{y}_{st})}{Var(\bar{y}_{SRS})} = \frac{Ms^2_{\bar{y}_i}}{s^2_{\bar{y}_{ij}}} \tag{9}$$

The simple sample variance of the 36 cans is

$$s^2_{\bar{y}_{ij}} = \frac{1}{36-1}\sum_{i=1}^{36}(\bar{y}_i - \bar{y}_{SRS})^2 \approx 7.38 \tag{10}$$

Thus, the design effect is

$$\frac{Ms^2_{\bar{y}_i}}{s^2_{\bar{y}_{ij}}} \approx \frac{3*4.53}{7.38} \approx 1.84 \tag{11}$$

(d) We have that the the intracluster correlation ρ, approximately satisfies $DEFF \approx 1 + (M-1)\rho$. Thus we can estimate ρ by solving

$$\rho \approx \frac{DEFF - 1}{M - 1} = \frac{1.84 - 1}{3 - 1} = 0.42. \tag{12}$$