36-303 Sampling, Surveys & Society
Homework 06 Solutions

April 13, 2012

1 Question 1

(a) To estimate the mean household and compute the standard error of our estimate as though from a SRS, we simply pretend as though the stratum labels are not there. We have that the mean is

$$\overline{y}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{25}[48,111 + 36,582 + ... + 78,404] \approx 61,856 \tag{1}$$

The standard error of the SRS estimator is

$$se(\overline{y}_{SRS}) = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{25}(y_i - \overline{y})^2} = \tag{2}$$

$$\frac{1}{\sqrt{25}}\sqrt{\frac{1}{24}[(48,111 - 68,156)^2 + ... + (78,404 - 68,156)^2]} \approx 2887 \tag{3}$$

(b) To compute post-stratum weights, we divide, for each stratum, the population proportion by the sample proportion, as shown below. The population proportion is given to us, while we compute the sample proportion for a stratum by dividing the number of households in that stratum by 25, the total sample size.

Household Size	Population Proportion	Sample Proportion	PS Weight
1	0.2575	0.12	2.15
2	0.3117	0.24	1.30
3	0.1750	0.28	0.62
4	0.1558	0.28	0.56
5	0.10	0.08	1.25

We then assign to each observation the relevant weight for its stratum, as shown below.

Household	y_i	Household Size	w_i
1	48111	1	2.15
2	36582	1	2.15
3	38245	1	2.15
4	46246	2	1.30
5	50309	2	1.30
6	58373	2	1.30
7	58800	2	1.30
8	58244	2	1.30
9	36898	2	1.30
10	63274	3	0.62
11	60458	3	0.62
12	73379	3	0.62
13	61650	3	0.62
14	73983	3	0.62
15	60696	3	0.62
16	62211	3	0.62
17	67884	4	0.56
18	61633	4	0.56
19	88080	4	0.56
20	69376	4	0.56
21	68719	4	0.56
22	55496	4	0.56
23	77201	4	0.56
24	92149	5	1.25
25	78404	5	1.25

Using these values for w_i and y_i, we calculate the post-stratification-weighted average household income thus:

$$\overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i} \approx 57,387$$

(c) **Taylor Series Method**. The estimate of $Var(\overline{y})$ using the Taylor Series Method is given by:

$$Var_{TS}(\overline{y}_w) \approx \frac{1}{(\sum_i w_i)^2}\left[Var(\sum_i w_i y_i) - 2\overline{y}_w Cov(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 Var(\sum_i w_i)\right] \quad (4)$$

where

$$Var(\sum_{i=1}^{n} w_i) \approx n \cdot \frac{1}{n-1} \sum_{i=1}^{n} (w_i - \overline{w})^2 \quad (5)$$

$$Var(\sum_{i=1}^{n} y_i w_i) \approx n \cdot \frac{1}{n-1} \sum_{i=1}^{n} (w_i y_i - \overline{wy})^2 \quad (6)$$

$$Cov(\sum_{i=1}^{n} y_i w_i, \sum_{i=1}^{n} w_i) \approx n \cdot \frac{1}{n-1} \sum_{i=1}^{n} (w_i y_i - \overline{wy})(w_i - \overline{w}) \quad (7)$$

Note that in the following, I will sometimes report numbers rounded. However, when I did the calculations I didn't round until the end – and neither should you if it's possible.

In order to use these formulas, we first need to calculate $\overline{w}$, $\overline{wy}$, and $\overline{y}_w$ using the values in the table on the preceding page. We have already calculated $\overline{y}_w = 57{,}387$ in part (b). We also have:

$$\overline{w} = \frac{1}{n} \sum_i w_i = 1 \quad (8)$$

$$\overline{wy} = \frac{1}{n} \sum_i w_i y_i = 57,387 \quad (9)$$

Plugging these values in equations (6) through (8), and using the w_i and y_i values from the table, we have

$$Var(\sum_{i=1}^{n} w_i) \approx 25 \cdot \frac{25}{24} \sum_{i=1}^{25} (w_i - 1)^2 \approx 7.25 \quad (10)$$

$$Var(\sum_{i=1}^{n} y_i w_i) \approx 25 \cdot \frac{1}{24} \sum_{i=1}^{25} (w_i y_i - 57,387)^2 \approx 14,578,434,918 \quad (11)$$

$$Cov(\sum_{i=1}^{n} y_i w_i, \sum_{i=1}^{n} w_i) \approx 25 \cdot \frac{1}{24} \sum_{i=1}^{25} (w_i y_i - 57,387)(w_i - 1) \approx 258,417 \quad (12)$$

Note that in order to use equation (4), we still must calculate $\left(\sum_i w_i\right)^2 = 625$.
Now, plugging into equation (4), we have:

$$Var_{TS}(\overline{y}_w) \approx \frac{1}{625} \left[14,578,434,918 - 2 \cdot 57,387 \cdot 258,417 + (57,387)^2 \cdot 7.25\right] \quad (13)$$

$$\approx 14,077,467 \quad (14)$$

Since this is the variance of the estimator, the standard error of the estimator is the square root of this, which is roughly 3752. Note that this is a fair amount higher than the standard error of the estimator based on a simple random sample.

2 Question 2: Jackknife

Using the jackknife method, we create 25 different replicate data sets, such that the i^{th} replicate data set contains all of the observations except the i^{th} one. Because leaving out one observation changes our sample proportions, it forces us to recalculate our weights. (Note that the population proportions stay the same!)

First recall the old characteristics of each stratum.

Household Size	Population Proportion	Old Sample Proportion	PS Weight
1	0.2575	0.12	2.15
2	0.3117	0.24	1.30
3	0.1750	0.28	0.62
4	0.1558	0.28	0.56
5	0.10	0.08	1.25

Below, the recalculated sample proportions.

	Replicates 1-3	Replicates 4-9	Replicates 10-16	Replicates 17-23	Replicates 24-25
1	0.08	0.13	0.13	0.13	0.13
2	0.25	0.21	0.25	0.25	0.25
3	0.29	0.29	0.25	0.29	0.29
4	0.29	0.29	0.29	0.25	0.29
5	0.08	0.08	0.08	0.08	0.04

And finally, the recalculated weights, obtained by dividing the population proportion for each stratum by its recalculated weight.

	Replicates 1-3	Replicates 4-9	Replicates 10-16	Replicates 17-23	Replicates 24-25
1	3.09	2.06	2.06	2.06	2.06
2	1.25	1.5	1.25	1.25	1.25
3	0.60	0.60	0.70	0.60	0.60
4	0.53	0.53	0.53	0.62	0.53
5	1.2	1.2	1.2	1.2	2.4

Given these new weights, our problem amounts to doing Part (b) of Question 1 25 times – i.e. assigning a weight to each of the (now 24) observations in a given replicate, and calculating $\frac{\sum_i w_i y_i}{\sum_i w_i}$ – to obtain a different estimate of weighted-average household income for each replicate data set. They are given below, rounded as usual.

Replicate Data Set	$\overline{y}_w^{(r)}$
1	56469
2	57953
3	57739
4	57713
5	57460
6	56957
7	56931
8	56965
9	58296
10	57440
11	57522
12	57145
13	57487
14	57128
15	57515
16	57471
17	57436
18	57598
19	56912
20	57397
21	57414
22	57758
23	57194
24	56700
25	58074

We can calculate the mean of these, which is the same as it was in part (b) of question 1:

$$\overline{y}_{JK} \;=\; \frac{1}{n}\sum_{r=1}^{n}\overline{y}_w^{(r)} \approx 57,387$$

Finally, we have from the handout that

$$Var_{JK}(\overline{y}_w) = \frac{n-1}{n}\sum_{r=1}^{n}(\overline{y}_w^{(r)} - \overline{y}_{jk})^2 = \frac{24}{25}\sum_{r=1}^{25}(\overline{y}_w^{(r)} - 57,387)^2 \approx 4,195,116$$

Taking the square root of that, the standard error of our estimator is roughly 2048. This is lower than that obtained by Taylor Series method, and also lower than the standard error for simple random sample! The latter is unusual behavior, as standard errors calculated using the jackknife and the Taylor Series approaches are usually but not always larger than those for a simple random sample.