STAT 3014/3914

Semester 2	Applied Statistics	2015

Tutorial 13

1. A company provides its salesmen with cars for work-related travelling. The company operates in 12 branches with 810 cars in all and a SRS of 4 branches is selected.

   (a) For each selected branch, the number of cars M_i and the total distance y_i (in thousands of miles) travelled last year by cars are reported in the following table:

Branch i	**No. of cars** M_i	**Total mileage** y_i
2	60	1459.2
5	110	3036.0
8	20	568.2
9	50	1277.5
Sum	240	6340.9

$$\overline{y} = 1585.225 \sum_i y_i^2 = 13301418.13 \sum_i M_i^2 = 18600 \sum_i M_i y_i = 496751 \, s_y^2 = 1083221.64$$

   Estimate the average number of miles travelled by a company car in the past year and its variance using both the naive and ratio estimators.

   (b) If a 2-stage sampling is performed to result in the following data:

Branch	M_i	m_i	**Sampled** y_{ij}	$\bar{y}_{i\cdot}$	s_{yi}^2	$\hat{y}_i = M_i \bar{y}_{i\cdot}$
1	60	10	30.40, 25.19, 25.71, 25.07, 24.18 24.56, 23.22, 25.42, 21.80, 25.27	25.082	4.90684	1504.92
2	110	10	29.26, 28.72, 25.94, 24.22, 29.41 27.24, 32.16, 33.26, 32.74, 22.92	28.587	12.55613	3144.57
3	20	10	23.69, 31.29, 31.21, 31.77, 28.12 28.99, 27.30, 28.15, 30.74, 27.76	28.902	6.09482	578.04
4	50	10	32.00, 24.11, 25.29, 22.79, 22.63 22.32, 25.69, 24.48, 27.00, 32.50	25.881	13.42161	1294.05
Sum	240	40				6521.58

$$\overline{\hat{y}} = 1,630.395 \quad \sum_{i=1}^{n} \hat{y}_i^2 = 14,161,800.33 \quad \sum_{i=1}^{n} \hat{y}_i M_i = 512,461.2 \quad \sum_{i=1}^{n} M_i^2 = 18,600$$

   Estimate the average number of miles travelled by a company car in the past year.

2. A garment manufacturer with 90 plants wants to estimate the average number of hours due to machine breakdown in the past months. He uses cluster sampling regarding each plant as a cluster of machines. Because of the large number of machines in each plant, he uses a two-stage cluster sampling by drawing $n = 10$ plants in stage 1 and subsamples approximately 20% of the machines in each plant.

   Using the data in the following table, estimate the average breakdown hour per machine using ratio estimator and naive estimator, and place a bound on the error of estimations. He has a record of totally 4500 machines in all plants. [4.5988, 0.4623; 4.8012, 0.3871]

Downtime for sewing machines

Plant	M_i	m_i	**Downtime (in hrs)**	$\overline{y}_i$	s_{yi}^2
1	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5	5.40	11.38
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5	4.00	10.67
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4	5.67	16.75
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10	4.80	13.29
5	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3	4.30	11.12
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4	3.83	14.88
7	42	8	3, 7, 6, 7, 8, 4, 3, 2	5.00	5.14
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3	3.85	4.31
9	40	8	6, 4, 7, 3, 9, 1, 4, 5	4.88	6.13
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4	5.00	11.80

3. Suppose that the hospital's purchases of a given pharmaceutical product Y and the size of the hospital as measured by the number of its beds (X) follows a linear relationship $Y \simeq RX$.

   (a) A random sample of $n = 4$ hospitals will be selected with PPS with probabilities of selection $p_i = \dfrac{x_i}{186,030}$ and information is given in the following table.

Hosp. i	**No. of beds** x_i	**Purchases of Y** y_i	**Purchases of Z** y_i'	**Assigned no.**	**Selection prob.** $p_i = \frac{x_i}{X}$	$\frac{y_i}{p_i}$	$\frac{y_i'}{p_i}$
1	675	500	1	000001-000675	$\frac{675}{186,030} = 0.0036$	144,690	275.60
2	450	350	0	000676-001125	$\frac{450}{186,030} = 0.0024$	137,800	0.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1,158	1,500	1,100	1	184531-186030	$\frac{1,500}{186,030} = 0.0081$	136,422	124.02
Total	186,030				1.0000		

   (i) Use the random numbers 001052, 185953, 000600, 000987 to select hospitals.

   (ii) Estimate the average purchases of Product Y per hospital and the proportion of hospital which purchase product Z using the HH estimator. Provide estimates of the variance for the estimators.

(b) For a sample $n = 3$ hospitals with IPPS and inclusion probabilities $\pi_i = \dfrac{3x_i}{186,030}$, the sample consists of hospitals 1, 2, and 1,158 with the information as given below:

Hospital i	**Number of beds** x_i	**Purchases of Y** y_i	$\pi_i = \frac{nx_i}{X}$
1	675	500	$\frac{3\times675}{186,030} = 0.0109$
2	450	350	$\frac{3\times450}{186,030} = 0.0073$
⋮	⋮	⋮	⋮
1,158	1,500	1,100	$\frac{3\times1,500}{186,030} = 0.0242$
Total	186,030		3

(i) Estimate the average purchases and proportion of purchases of product Y per hospital using the HT estimator.

(ii) Suppose that the 3 hospitals obtained in (a) were actually drawn by systematic PPS sampling. Estimate the average purchases by hospitals and give an approximate of its standard error.

4. A three-stage cluster sampling is conducted without replacement and the procedures includes

(a) Select l of the L districts.

(b) Select m_i of the M_i households in selected district l and

(c) Select n_{ij} of the N_{ij} members of selected household j in selected district i.

The following table lists the population.

Three-stage sampling

District	**Household**	**Household Members**	**Age**	**Stage 1** $\frac{l}{L}$	**stage 2** $\frac{m_i}{M_i}$	**Stage 3** $\frac{n_{ij}}{N_{ij}}$	**Inclusion prob.** $\pi_i = \frac{l}{L}\frac{m_i}{M_i}\frac{n_{ij}}{N_{ij}}$
1. North	1. Brown	1. John		1/2	2/2	1/2	1/4
1. North	1. Brown	2. Mary		1/2	2/2	1/2	1/4
1. North	2. Smith	1. Jane		1/2	2/2	1/1	1/2
2. South	1. Jones	1. Alice		1/2	2/3	1/2	1/6
2. South	1. Jones	2. Peter $\surd$	25	1/2	2/3	1/2	1/6
2. South	2. Cox	1. Jacob		1/2	2/3	1/3	1/9
2. South	2. Cox	2. Sarah		1/2	2/3	1/3	1/9
2. South	2. Cox	3. Melinda		1/2	2/3	1/3	1/9
2. South	3. Elton	1. David $\surd$	60	1/2	2/3	1/1	1/3
							2

Based on this sampling method, Peter Jones and David Elton are selected and their ages are 25 and 60 respectively. Estimate the population mean age using the HT estimator and report its standard error.

Extra exercise

1. 315 primary schools in a city were grouped into 30 school districts with each school district containing approximately 10 schools. A simple random sample of 3 school districts was taken and then within each sample school district, a simple random sample of 4 schools was taken for purposes of estimating the number of school children in the city that are colour-blind. The data shown in the table were obtained from this sample.

Sample district, i	**No. of schools,** M_i	**No. of sample schools,** m_i	**No. of colour-blind children,** y_{ij}	**Mean** $\overline{\overline{y}}_i$	**Var.** s_i^2
1	10	4	1,3,3,4	2.75	1.5833
2	12	4	3,4,0,1	2.00	3.3333
3	9	4	4,2,0,1	1.75	2.1875

Estimate the total number of colour-blind school children in the city and obtain a standard error for the estimated total using both *the ratio estimator method* and *the number-raised/ordinary estimator method*. Compare the results. Which estimator would you prefer? Explain why. [683.3468, 91.1903; 672.5, 105.4061]

2. A coaching school wants to estimate the average score of its students in a public examination. Students in the school are divided into classes of approximately 10 students. Cluster sampling is adopted with classes being the sampling unit in the first stage. Owing to the available resources, the sample will contain approximately 40 students from 41 classes with a total enrollment of 426 students.

   (a) A teacher uses a 1-stage cluster sampling and selects 4 classes randomly. The first 4 columns of the following table give the informaton of the 1-stage cluster sample.

Class i	**Class size** M_i	**Student scores** (a) y_{ij}	**Class total** (a) y_i	**Sample total** (b) $\sum_{j\in\mathcal{S}'} y_{ij}$	**Sample variance** (b) s_{yi}^2	**Estimated class total** (b) $\hat{y}_i$
1	10	7,8,6,8,4,5,8,9,10,7	72	35	5.00	
2	11	8,6,9,10,10,8,9,6,7,8,7	88	50	1.47	
3	9	7,10,6,3,4,5,8,7,9	59	34	3.70	
4	10	7,5,8,6,9,5,7,7,2,4	60	33	7.30	

Some summary statistics are given below.

$$\sum_i y_i = 279;\ \sum_i M_i = 40;\ \sum_i y_i^2 = 20,009;\ \sum_i M_i y_i = 2,819;\ \sum_i M_i^2 = 402.$$

   (i) Estimate the average score per student using the *ordinary* estimate and provide an estimate of its standard error.

   (ii) Would you expect the *ordinary* estimate to be larger, more or less the same, or smaller than the ratio estimate? Would you prefer ratio estimator or ordinary estimator in this situation? Explain.

(b) Another teacher suggests to use a 2-stage cluster sampling. She randomly selects 8 classes and finds that the 4 classes selected by the officer are also included. The 4 additional classes selected are given in the following table. The scores in each class are listed in alphabetical order of the students' surnames and a systematic sample that includes the 1st, the 3rd, the 5th and so on, of the scores is drawn from each class.

Class i	**Class size** M_i	**Student marks** (a) y_{ij}	**Class total** (a) y_i	**Sample total** (b) $\sum_{j\in\mathcal{S}'} y_{ij}$	**Sample variance** (b) s_{yi}^2	**Estimated class total** (b) $\hat{y}_i$
5	9	6,7,7,5,4,3,5,6,4	47	26	1.70	
6	11	5,3,2,4,5,3,4,2,1,5,0	34	17	4.57	
7	11	4,8,6,7,4,5,3,6,9,8,2	62	28	6.27	
8	9	8,9,7,8,5,7,7,4,9	64	36	2.20	

(i) Combine the two tables and complete the last column in your answer book.

(ii) Assuming that the systematic sample from each class is equivalent to a simple random sample, estimate the average score per student using the *naive* estimate and provide an estimate of its standard error. To simplify your calculation, use $(1-\frac{m_i}{M_i}) \simeq (1-\frac{5}{10}) = \frac{1}{2}$ to be the approximate finite population correction factor for all classes.

(iii) Consider the scores in classes 1-4 and classes 5-8 separately. Would respectively the mean estimate and its standard error estimate in part (i) using 1-stage cluster sampling be under-estimated, correct or over-estimated? What type of error, sampling error or non-sampling error, will these two estimates subject to?

(iv) Would you prefer 1-stage or 2-stage cluster sampling in this situation? Explain briefly.

3. (a) Given the total sample size of elements, state with reasons why you would choose between 1-stage or 2-stage cluster sampling in the following situations:

(i) when there is high variability in cluster sizes,

(ii) when the variability within each cluster is low but the variability between clusters is high.

(b) A cable TV company wants to perform a survey on the average number of hours that the residents of a certain district watch the cable TV per day. There are totally $N = 48$ blocks with $M = 5200$ households in the district. A random sample of $n = 6$ blocks are selected and the total number of households in each block and the total number of hours that these households watch the cable TV in each block are recorded and listed below:

Block i	**Number of household** M_i	**Number of hours watching cable TV** y_i	**Average per household** $\overline{y}_i = y_i/M_i$
1	161	334	
2	148	356	
3	83	245	
4	157	412	
5	96	207	
6	103	315	

(i) Complete the last column of the table which gives the average number of hours watching cable TV per household for the 6 blocks selected. Consider a new estimate $\overline{\overline{Y}}_1$ which use the sample mean of the 6 averages $\overline{y}_i = y_i/M_i$ as the estimate for the overall average number of hours watching cable TV per household:

$$\overline{\overline{Y}}_1 = \frac{\sum_{i=1}^{6} \overline{y}_i}{n}.$$

with

$$\text{var}(\overline{\overline{Y}}_1) = (1 - \frac{n}{N})\frac{s^2_{\overline{y}}}{n}$$

where

$$s^2_{\overline{y}} = \frac{\sum_i \overline{y}_i^2 - \frac{(\sum_i \overline{y}_i)^2}{n}}{n-1}.$$

Estimate the average number of hours watching cable TV per household in the estate and provide an estimate of the standard error for the estimator.

(ii) Estimate the average number of hours watching cable TV per household in the estate using the *naive* estimator, $\overline{\overline{Y}}_2$ and provide an estimate of the standard error for the estimator. You may find the following summary statistics useful:

$$\sum_i y_i = 1,869;\ \sum_i M_i = 748;\ \sum_i y_i^2 = 610,135;\ \sum_i M_i y_i = 243,798; \sum_i M_i^2 = 99,188.$$

(iii) Estimate the average number of hours watching cable TV per household in the estate using the *ratio* estimator, $\overline{\overline{Y}}_3$, and provide an estimate of the standard error for the estimator.

(iv) State which estimator, $\overline{\overline{Y}}_1$, $\overline{\overline{Y}}_2$ or $\overline{\overline{Y}}_3$ you think is suitable in each of the following situations. Explain briefly.

(1) M is unknown, the variability of M_i is high and M_i is highly and positively correlated with y_i.

(2) M is known and the variability of M_i is low.

4. Give conditions when 1-stage cluster sampling should be preferred to 2-stage cluster sampling and when 2-stage cluster sampling should be preferred to 1-stage cluster sampling.