Semester 2	Applied Statistics	2015

Assignment for Sample Survey

1. In the election of president and vice-president, there are 3 candidates, C_1, C_2 and C_3 and each voter is allowed to select up to 2 candidates (that is, the voter can either select 1 or 2 candidates). Let N_{ij} ($i \neq j,\ i, j = 1, 2, 3$) be the total number of voters who vote for C_i and C_j in the election, and N_{ii} ($i = 1, 2, 3$) be the total number of voters who vote for C_i only. Suppose there is a total of N voters in the election and let P_i ($i = 1, 2, 3$) be the actual proportion of voters who vote for C_i in the election. Note that $N = \sum_{i=1}^{3} \sum_{j=i}^{3} N_{ij} = N_{11} + N_{12} + N_{13} + N_{22} + N_{23} + N_{33}$ and $N_{ij} = N_{ji}$ for any i and j.

   (a) Express P_i ($i = 1, 2, 3$) and hence $P_1 - P_2$ in terms of the N_{ij}'s and N. Say $P_1 = (N_{11} + N_{12} + N_{13})/N$.

   (b) Suppose a simple random sample of $n = 200$ voters is drawn on the election day. The number of voters n_{ij} ($i \leq j,\ i, j = 1, 2, 3$) in the sample who vote for C_i and C_j were given in the following table. (Note that n_{ii} denotes the number of voters in the sample who vote for C_i only.)

		j = 1	2	3
i	1	35	72	23
	2		26	16
	3			28

   Estimate $P_1 - P_2$, the difference in the proportion of voters who vote for C_1 and C_2. Provide a 95% confidence interval. Assume that N is large enough to neglect the finite population correction and that the estimator is normally distributed. [Ans: 0.08,(-0.017,0.177)]

   Hint: Define, for $i = 1, 2, \cdots, N$,

$$y_i = \begin{cases} 1, & \text{if the } i\text{th voter votes for } C_1 \text{ but not for } C_2, \\ -1, & \text{if the } i\text{th voter votes for } C_2 \text{ but not for } C_1, \\ 0, & \text{otherwise.} \end{cases}$$

2. A simple random sample of 300 employees was chosen from the population of a large company with 4800 employees. Each selected employee was asked whether he owned or rented his accommodation. Their monthly income y was also recorded. Results are as follows:

	Number of persons (n_i)	**Average monthly income ($\bar{y}$)**	**Sample standard deviation of monthly income (s)**
Owning (C_1)	100	\$12000	\$400
Renting (C_2)	200	\$8000	\$100

   It is known that about 30% of employees own their accommodation.

(a) Estimate the average monthly income for all employees of the company and its standard error. [9200; 13.1588]

(b) Estimate the average monthly income and its standard error among those employees who owned their accommodation using the estimator

$$\hat{\bar{Y}}_{pst,1m1} = N\bar{y}'/N_1$$

where $y_i' = y_i$ if $y_i \in C_1$ and 0 otherwise. [13333.33; 1056.7245]

(c) Estimate the average monthly income and its standard error among those employees who owned their accommodation using the estimator

$$\hat{\bar{Y}}_{pst,1m2} = n\bar{y}'/n_1 = \bar{y}_1.$$

[12000; 40.82483]

(d) Suppose the proportion that 30% of employee owned their accommodation is unknown and double sampling is used. 10% of the sampled employees as listed in the table are asked for the information of monthly income. The sample mean and sample standard deviation for the sub-samples are the same as those given in the table. Estimate the average monthly income for all employees of the company and its standard error. You may ignore fpc in the calculation. Compare the post-stratification estimator in (a) with this double sampling estimator. [9333.333; 117.694]

3. A transport officer wants to estimate the average weekly travel expenses for people living in a certain district. He obtains a list of 100 households in the district and samples 8 households from the list. The household sizes of the sampled households vary from 1 to 6.

(a) It is decided that for those households of sizes 1, 2 or 3, all members in the household will be surveyed. For those households of sizes 4 or more, a subsample of size 3 will be taken. The following table shows the result:

House-hold i	**Size** M_i	**Sample size** m_i	**Weekly travel expense** y_{ij}	**Sample mean** $\bar{y}_{i\cdot}$	**Sample var.** s^2_{yi}	**Total** y_i / **Est. total** $\hat{y}_i$
1	1	1	42.5	42.5	N.A.	42.5
2	1	1	7.0	7.0	N.A.	7.0
3	2	2	12.2 46.0	29.1	N.A.	58.2
4	3	3	0.0 35.4 46.8	27.4	N.A.	82.2
5	4	3	5.8 30.2 40.8	25.6	322.12	
6	4	3	18.2 25.6 42.6	28.8	156.52	
7	5	3	14.6 22.4 25.4	20.8	31.08	
8	6	3	0.0 34.6 41.0	25.2	486.52	

Complete the four missing cells of estimated totals $\hat{y}_i$ for households $i = 5, \ldots, 8$.

(b) To estimate the average weekly travel expenses for people living in the district, which estimator should he use? Explain Why. Calculate the estimate and provide the variance estimate of the estimator he uses.

(Hint: For household $i, i = 1, \ldots, 4$, you may assume the additional variabilities due to y_i are zero since y_i are fixed and the sampling fraction $\frac{m_i}{M_i}$ are one.) [25.4885, 2.6799]

(c) Suppose the transport officer also wants to estimate the average weekly travel expenses for households in the district, which estimator should he use? Calculate the estimate and provide the variance estimate of the estimator he uses. [82.8375, 243.4009]

4. (Optional with bonus) The following table shows the five farms in a district, their number of farm acres x planted in trees (as an auxiliary variable) and y the monthly income (in thousand dollars) from fruits selling.

Unit i	**Farm acres** x_i	**Monthly income** (\$000) y_i
1	20.7	63
2	5.2	16
3	9.6	30
4	10.3	31
5	14.2	46

(a) Using the following random numbers:

0.899 0.915 0.201

select a PPS sample of 3 units with replacement. Show the assigned numbers, the random numbers used, and the selected units. Compute the Hansen-Hurwitz (HH) estimate of the population mean $\overline{Y}$ and its standard error estimate. [38.08941; 0.78384]

(b) Repeat part (a) for a IPPS sample of 3 units using the Madow 1949 method. List all possible samples for different random numbers. Using the random number 0.899 to select the sample, estimate the population mean $\overline{Y}$ using the Horvitz-Thompson (HT) estimator. Standard error estimate is NOT required. [37.1705]