STAT 3014/3914

| Semester 2 | Applied Statistics | 2015 |
|---|---|---|

**Tutorial 12**

1. A certain manufacturing firm produces a product that is packaged under two brand names, for marketing purpose. These two brands serve as strata for estimating potential sales volume for the next quarter. A simple random sample of customers is contacted and asked to provide a potential sales figure $y$ (in number of units) for the coming quarter and the sales figure $x$ for the previous quarter for the brand they chose. The data are given in the following table.

| **Brand I** | | **Brand II** | |
|---|---|---|---|
| $x_i$ | $y_i$ | $x_i$ | $y_i$ |
| 204 | 210 | 137 | 150 |
| 143 | 160 | 189 | 200 |
| 82 | 75 | 119 | 125 |
| 256 | 280 | 63 | 60 |
| 275 | 300 | 103 | 110 |
| 198 | 190 | 107 | 100 |
| | | 159 | 180 |
| | | 63 | 75 |
| | | 87 | 90 |

   The sample was taken from a list of 300 customers. The total sales in the previous quarter was 24500 units for brand I and 21200 units for brand II.

   (a) Estimate the total sales using the *Hartley Ross* estimator.

   (b) Estimate the rate of increase of sales of Brand I product and its standard error.

   (c) Estimate the total sales of Brand I product and its standard error.

2. Manufacturing companies A and B with respectively $N_A = 1,000$ and $N_B = 1,500$ employees, wishes to estimate the average man-hours lost in current year due to sickness and they use the number of man-hours lost due to sickness last year as an auxiliary variable. A preliminary study of $n = 10$ employee records is made in both companies and the results are given in the following table. Records show that the total number of man-hours lost because of sickness for the previous year are $X_A = 16,300$ and $X_B = 12,800$ respectively for companies A and B. Assume that companies A and B together form the population of workers of interest in this problem.

(a) Find the *separate* ratio estimate of the average man-hours lost in current year due to sickness $\overline{Y}$. The data for the two companies are given below:

| | $x_{Ai}$ | $y_{Ai}$ | $y_{Ai} - r_A x_{Ai}$ | $x_{Bi}$ | $y_{Bi}$ | $y_{Bi} - r_B x_{Bi}$ |
|---|---|---|---|---|---|---|
| 1 | 12 | 13 | 0.39326 | 10 | 8 | 2.10256 |
| 2 | 24 | 25 | -0.21349 | 8 | 0 | -4.71795 |
| 3 | 15 | 15 | -0.75843 | 0 | 4 | 4.00000 |
| 4 | 30 | 32 | 0.48314 | 14 | 6 | -2.25641 |
| 5 | 32 | 36 | 2.38202 | 12 | 10 | 2.92308 |
| 6 | 26 | 24 | -3.31461 | 6 | 0 | -3.53846 |
| 7 | 10 | 12 | 1.49438 | 4 | 2 | -0.35897 |
| 8 | 15 | 16 | 0.24157 | 0 | 4 | 4.00000 |
| 9 | 0 | 2 | 2.00000 | 8 | 4 | -0.71795 |
| 10 | 14 | 12 | -2.70787 | 16 | 8 | -1.43590 |

| | $n$ | Mean | Cov | SD |
|---|---|---|---|---|
| $x_A$ | 10 | 17.80 | | 9.99 |
| $y_A$ | 10 | 18.70 | 101.822 | 10.36 |
| $y_A - r_A x_A$ | 10 | -0.000 | | 1.86 |
| $x_B$ | 10 | 7.80 | | 5.45 |
| $y_B$ | 10 | 4.60 | 10.356 | 3.41 |
| $y_B - r_B x_A$ | 10 | 0.000 | | 3.12 |

Note that the columns of $y_{Ai} - r_A x_{Ai}$ and $y_{Bi} - r_B x_{Bi}$ are given to show that they sum to zero.

(b) Find the *combinated* ratio estimate of the average man-hours lost in current year due to sickness $\overline{Y}$.

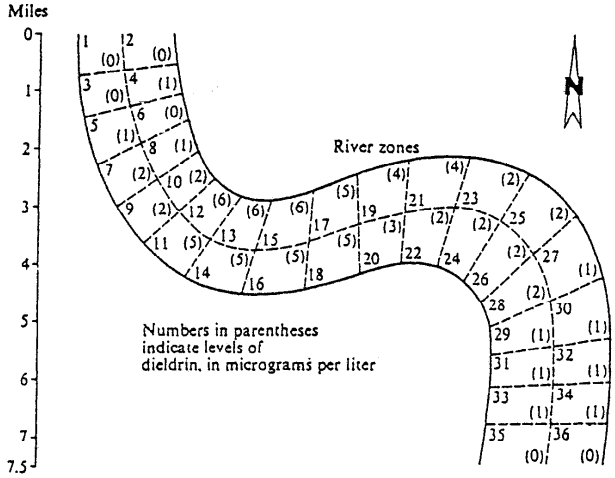
| | $x_{Ai}$ | $y_{Ai}$ | $y_{Ai} - r_C x_{Ai}$ | $x_{Bi}$ | $y_{Bi}$ | $y_{Bi} - r_C x_{Bi}$ |
|---|---|---|---|---|---|---|
| 1 | 12 | 13 | 2.58400 | 10 | 8 | -0.68000 |
| 2 | 24 | 25 | 4.16800 | 8 | 0 | -6.94400 |
| 3 | 15 | 15 | 1.98000 | 0 | 4 | 4.00000 |
| 4 | 30 | 32 | 5.96000 | 14 | 6 | -6.15200 |
| 5 | 32 | 36 | 8.22400 | 12 | 10 | -0.41600 |
| 6 | 26 | 24 | 1.43200 | 6 | 0 | -5.20800 |
| 7 | 10 | 12 | 3.32000 | 4 | 2 | -1.47200 |
| 8 | 15 | 16 | 2.98000 | 0 | 4 | 4.00000 |
| 9 | 0 | 2 | 2.00000 | 8 | 4 | -2.94400 |
| 10 | 14 | 12 | -0.15200 | 16 | 8 | -5.88800 |

| | $n$ | Mean | Cov. | SD |
|---|---|---|---|---|
| $x_A$ | 10 | 17.80 | | 9.99 |
| $y_A$ | 10 | 18.70 | 101.822 | 10.36 |
| $y_A - r_C x_A$ | 10 | 3.25 | | 2.39 |
| $x_B$ | 10 | 7.80 | | 5.45 |
| $y_B$ | 10 | 4.60 | 10.356 | 3.41 |
| $y_B - r_C x_B$ | 10 | -2.17 | | 4.00 |

Note that the columns of $y_{Ai} - r_C x_{Ai}$ and $y_{Bi} - r_C x_{Bi}$ are given to show that they do NOT sum to zero.

(c) Discuss the situations when the separate ratio estimator or combined ratio estimator should be used.

3. Suppose that a study is planned of the level of the pesticide dieldrin, which is believed to be a carcinogen, in a 7.5 mile stretch of a particular river. To assure the representativeness, a map of the river is divided into 36 zones, and 9 systematic samples of these zones are formed. Water samples will be drawn by taking a boat out to the geographic center of the designated zone, and drawing a grab sample of water from a depth of several centimeters below the surface level. The levels of dieldrin, in micrograms per liter, for each of these zones are shown in parentheses.




(a) Calculate the sample mean for each of the 9 possible systematic samples and the *true variance* of the mean estimator.

(b) What advantages can you identify for this method of sampling the river over simple random sampling? Show that systematic sampling are more efficient than simple random sampling.

(c) Suppose 3 systematic samples with random starts of 2, 4 and 8 are selected, estimate the average level of dieldrin and its variance for this stretch of the river.

## Extra exercise

1. A property company has 3 branches in a district. The company manager wants to study the total amount of commissions received by all agents, $Y$ in thousands during a particular month. He obtains a simple random sample of 15 agents out of 37 agents and then he stratifies them according to certain auxiliary information: branches simply for convenience purpose or 'the length of service in the company' which will give to homogenous strata. The following is the data that is stratified according to these two criteria:

| **Stratified according to branches** | | | | |
|---|---|---|---|---|
| **Branch** $i$ | **Stratum size** $N_i$ | **Commission received** $y_{ij}$ | **Sample mean** $\overline{y}_i$ | **Sample variance** $s^2_{yi}$ |
| 1 | 12 | 18.0 52.0 65.2 78.5 93.0 | 61.34 | 819.328 |
| 2 | 10 | 25.5 32.5 56.5 70.5 96.5 | 56.30 | 833.200 |
| 3 | 15 | 29.7 30.0 48.5 83.0 88.3 | 55.90 | 799.045 |
| **Stratified according to length of service in the company** | | | | |
| **Length of service** $i$ | **Stratum size** $N_i$ | **Commission received** $y_{ij}$ | **Sample mean** $\overline{y}_i$ | **Sample variance** $s^2_{yi}$ |
| 1-2 yrs | 17 | 18.0 25.5 29.7 30.0 32.5 52.0 | 31.28 | 128.942 |
| 3-5 yrs | 12 | 48.5 56.5 65.2 70.5 83.0 | 64.74 | 174.613 |
| 6+ yrs | 8 | 78.5 88.3 93.0 96.5 | 89.08 | 60.989 |

(a) Estimate the total commission received for all 37 agents in the 3 branches when the data is stratified according to branches and provide an estimate of standard error for this estimator. [2137.58; 232.762]

(b) Estimate the total commission received for agents in the 3 branches when the data is stratified according to 'the length of stay in the company' and provide an estimate of standard error for this estimator. [2021.28; 91.851]

(c) Compare the variance estimates in part (a) and (b), which auxiliary information leads to more efficient estimate? Explain briefly.

(d) Estimate the total commission received for agents in the 3 branches under *simple random sampling* and provide an estimate of standard error for this estimator. Compare this standard error estimate with that in part (a) and (b), does stratification improve the efficiency of the estimator? [2140.327; 195.876]

(e) Assuming that the sample size $n$ is sufficiently large and the sampling fraction $\frac{n}{N}$ is sufficiently small, show that

$$\text{var}(\widehat{Y}_{srs}) - \text{var}(\widehat{Y}_{pst}) \simeq \frac{N^2}{n}(s^2 - \sum_l W_l s_l^2) = \frac{N^2}{n}\Delta s^2$$

where $s^2$ is the sample variance for the simple random sample. Calculate $\Delta s^2$ for the two stratification methods, according to branches and 'the length of service in the company' in part (a) and (b) respectively. Comment on the difference in values. [-107.8389, 577.954]

2. A government official wants to estimate the total annual profit last year for all 9,650 trading firms in a certain industry employing a population of 72,730 people. Two methods were suggested:

Method A: Conduct a stratified simple random sampling of 100 trading firms ($n = 100$) stratified according to their number of employee and use proportional allocation to locate firms into strata.

Method B: Take a simple random sample of 100 trading firms and use their number of employee as an auxiliary variable in the ratio estimation.

(a) Both methods involve the use of auxiliary information (on number of employee). Under what condition will each of the above two methods be more favorable?

(b) When method A is applied, the following data shows the stratification of all the trading firms in that industry by the number of people employed together with the stratum size, average annual profit in ten thousands stratum variance in each stratum.

| **Firm size (No. of employee)** | **Number of firm** $N_l$ | **Average annual profit** $\overline{Y}_l$ | **Standard deviation** $S_l$ |
|---|---|---|---|
| 0- 5 | 2940 | 7.4 | 3.2 |
| 6-10 | 3530 | 16.3 | 6.3 |
| 11-20 | 2110 | 24.3 | 10.1 |
| 21+ | 1070 | 42.2 | 16.5 |

(i) For a sample of 100 firms, compute the sample sizes in each stratum under *Neyman allocation*. [13, 32, 30, 25]

(ii) Using the Neyman allocation in part (i), a stratified sample of 100 firms is obtained and shown in the following table. Estimate the total annual profit of last year for all the trading firms in that industry and provide a 95% confidence interval for the estimate. [170,214; (153,888.56, 186,539.460)]

| **Firm size** $l$ | **Sample mean of annual profit** $\overline{y}_l$ | **Sample variance** $s_l^2$ |
|---|---|---|
| 0- 5 | 8.7 | 16.1604 |
| 6-10 | 15.2 | 67.6996 |
| 11-20 | 20.4 | 151.5361 |
| 21+ | 44.8 | 232.2576 |

(c) When method B was applied, with

$$\sum_i x_i = 825,\ \sum_i y_i = 1926.5,\ \sum_i x_i^2 = 8674,\ \sum_i x_i y_i = 25707,\ \sum_i y_i^2 = 82671,$$

estimate the total annual profit of last year for all the trading firms in that industry using *ratio* estimation and provide the standard error of this estimator. [169835.57; 9604.842]

(d) Base on the results obtained in (b) and (c), state which method you would prefer and explain why.

3. A simple random sample of 300 employees was chosen from a large factory. Each selected employee was asked whether he owned or rented his accommodation. His quarterly income $y$ was also recorded. Results are as follows:

| | **Number of persons ($n_i$)** | **Average income ($\overline{y}_i$)** | **Sample standard deviation of income ($s_i$)** |
|---|---|---|---|
| Owning ($i$=1) | 100 | \$12,000 | \$400 |
| Renting ($i$=2) | 200 | \$8,000 | \$100 |

(a) Estimate the average quarterly income of employees of the factory and provide a standard error.

(b) Suppose that only 10% of the sample as listed in the table are asked for the information of quarterly income. The sample mean and sample standard deviation for the sub-samples are the same as those given in the table. Estimate the overall mean quarterly income of employees of the factory and its standard error. [9,333.333; 117.694]

4. A juice drink company wants to estimate the total current inventory of its product being held by its $N = 19,400$ dealers. The company wants to stratify the population of dealers into three groups according to their inventory levels for the previous year.

(a) The inventory levels for the previous year can only be obtained after interviewing the dealers. A simple random sample of $n' = 970$ dealers is drawn and the data are then stratified as below (in thousand dollars):

| **Stratum** $i$ | **Inventory last year** $x_i$ | **No. of dealers in the sample** $n_i'$ | **Sample mean** $\bar{x}'$ | **Sample variance** $s_{x,i}^{'2}$ |
|---|---|---|---|---|
| 1. Small | 0 - 0.999 | 272 | 8.5 | 16.1 |
| 2. Medium | 1 - 1.999 | 510 | 17.2 | 25.2 |
| 3. Large | >1.999 | 188 | 27.0 | 55.5 |

(i) Estimate the total inventory $X$ in thousands of dollars for all dealers and its *variance* using weights $W_l$ 0.30, 0.55 and 0.15 respectively for the three strata. [311,564; 9,985,643]

(ii) Estimate the population size $N_i$ for each stratum. [5,440; 10,200; 3,760]

(iii) Show that the Neyman allocation $n_i$ for a sample of size $n = 97$ using the $N_i$ determined in part (ii) and $s_{x,i}^{'2}$ as given in the table are 21, 49 and 27 respectively. [21; 49; 27]

(b) Subsamples of sizes $n_i$ determined in part (iii) are drawn from each post-stratified sample. The current inventory levels in thousand dollars for these dealers after stock checks are given below:

| **Stratum** $i$ | **Inventory last year** $x_i$ | **No. of dealers in the first sample** $n_i'$ | **Sub-sample size** $n_i$ | **Sub-sample mean** $\bar{y}_i$ | **Sub-sample variance** $s_{y,i}^2$ |
|---|---|---|---|---|---|
| 1. Small | 0 - 0.999 | 272 | 21 | 9.7 | 15.3 |
| 2. Medium | 1 - 1.999 | 510 | 49 | 18.3 | 28.5 |
| 3. Large | >1.999 | 188 | 27 | 31.5 | 60.2 |

(i) State and explain briefly the condition under which double sampling for stratification is preferred.

(ii) Estimate the total current inventory $Y$ in thousands of dollars for all dealers and its *variance estimate*. [357,868; 134,737,058.5]