1 mark There is 95% confidence that the true population total number of unit in class C is between 142 and 228

1. (a) In a simple random sample of size 100, from a population of size 500, there are 37 units in class C. Find the 95% confidence limits for the proportion and for the total number in class C in the population. **(10 marks)**

$n = 100$
$N = 500$
$a = 37$

$\hat{A} = Np$
$\hat{A} = 500 \times 0.37$
$\hat{A} = 185$

$p = \frac{a}{n} = \frac{37}{100} = 0.37$
$q = 1 - 0.37 = 0.63$

95% confidence for the proportion

$p \pm Z_{1-0.05/2} \sqrt{\frac{N-n}{N} \frac{pq}{n-1}}$

$0.37 \pm 1.96 \sqrt{\frac{500-100}{500} \times \frac{0.37 \times 0.63}{100-1}}$

0.37 ± 0.085

$(0.2849, 0.4551)$

4 marks

We have 95% confidence that the true population proportion of units in class C is between 0.2849 and 0.4551. (1 mark)

95% confidence interval for population total

$\hat{A} \pm Z_{\alpha/2} \sqrt{N(N-n) \frac{pq}{n-1}}$

$185 \pm 1.96 \sqrt{500(500-100) \frac{0.37 \times 0.63}{100-1}}$

$(142.47, 227.53)$ (4 mark)

(b) A simple random sample of 30 households was drawn from a city area containing 14848 households. The numbers of persons per household in the sample were as follows.

5,6,3,3,2,3,3,3,4,4,3,2,7,4,3,5,4,4,3,3,4,3,3,1,2,4,3,4,2,4

Estimate the total number of people in the area and compute the probability that this estimate is within ±10% of the true value. **(10 marks)**

$\bar{y} = \frac{\sum_{i=1}^{30} x_i}{30} = 3.467$ 1 mark

$s^2 = 1.4989$ 1 mark

(total number of people) $\hat{Y} = N\bar{y} = 14848 \times 3.467$ 1 mark

$\hat{Y} = 51473$

$P(|\hat{Y} - Y| < 0.1\hat{Y})$

$P\left(\frac{|\hat{Y} - Y|}{S.E(\hat{Y})} < \frac{0.1\hat{Y}}{S.E(\hat{Y})}\right)$

$P\left(|Z| < \frac{0.1 \times 51473}{S.E(\hat{Y})}\right)$

2 marks

Note $S.E(\hat{Y}) = N\sqrt{\frac{(1-f)s^2}{n}}$

$= 14848\sqrt{\frac{(1 - \frac{30}{14848}) \times 1.4989}{30}}$

$= 3315.54$

2 marks

Therefore

$P\left(|Z| < \frac{0.1 \times 51473}{3315.54}\right)$

$P(|Z| < 1.5525)$

$P(-1.5525 < Z < 1.5525)$

$P(Z < 1.5525) - P(Z < -1.5525)$

$0.9394 - 0.0606$

0.8788 (3 marks)

∴ The probability that the estimate is within ±10% of true value is 0.8788

2. (a) Show that the variance of the estimate, $\bar{y}_{st}$ is

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{(N_h - n_h)S_h^2}{N_h n_h})$$

where $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$ and $V(y) = \frac{(N-n)S^2}{Nn}$. **(10 marks)**

3 marks

$$V(\bar{y}_{st}) = V\left(\sum_{h=1}^{H} W_h \bar{y}_h\right) = V(W_1\bar{y}_1 + W_2\bar{y}_2 + \cdots + W_H\bar{y}_H)$$
$$= V(W_1\bar{y}_1) + V(W_2\bar{y}_2) + \cdots + V(W_H\bar{y}_H)$$

(2 marks) $\text{cov}(W_{h-1}\bar{y}_{h-1}, W_y\bar{y}_h) = 0$ since each stratum is independent

5 marks

$$\therefore V(\bar{y}_{st}) = W_1^2 V(\bar{y}_1) + W_2^2 V(\bar{y}_2) + \cdots + W_H^2 V(\bar{y}_H)$$
$$= W_1^2 \frac{N_1 - n_1}{N_1 n_1} S_1^2 + W_2^2 \frac{N_2 - n_2}{N_2 n_2} S_2^2 + \cdots + W_H^2 \frac{N_H - n_H}{N_H n_H} S_H^2 = \sum_{h=1}^{H} W_h^2 \frac{N_h - n_h}{N_h n_h} S_h^2$$

(b) Hayes (2000) took a stratified sample of New York City food stores. The sampling frame consisted of 1408 food stores with at least 4000 square feet of retail space. The population of stores was stratified into three strata using median household income within the zip code. The prices of a "market basket" of goods were determined for each store; the goal of the survey was to investigate whether prices differ among the three strata. Hayes used the logarithm of total price for the basket as the response y. Results are given in the following table:

Stratum, h	N_h	n_h	$\bar{y}_h$	s_h
1 Low income	190	21	3.925	0.037
2 Middle income	407	14	3.938	0.052
3 Upper income	811	22	3.942	0.070

Estimate $\bar{y}_{st}$ for these data and give a 95% confidence interval. **(10 marks)**

$$\bar{y}_{st} = \sum_{h=1}^{3} W_h \bar{y}_h = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3$$
$$= \frac{N_1}{N}\bar{y}_1 + \frac{N_2}{N}\bar{y}_2 + \frac{N_3}{N}\bar{y}_3$$
$$= \frac{190}{1408} \times 3.925 + \frac{407}{1408} \times 3.938 + \frac{811}{1408} \times 3.942$$
$$= 3.9386$$ (3 marks)

4 marks

$$V(\bar{y}_{st}) = \sum_{h=1}^{3} W_h^2 \frac{N_h - n_h}{N_h n_h} s_h^2$$
$$= \left(\frac{190}{1408}\right)^2 \frac{190-21}{190 \times 21} \times 0.037^2 + \left(\frac{407}{1408}\right)^2 \frac{407-14}{407 \times 14} \times 0.052^2 +$$
$$\left(\frac{811}{1408}\right)^2 \times \frac{811-22}{811 \times 22} \times 0.070^2 = 0.00008852$$

$\therefore$ 95% C.I

3 marks

$$\bar{y}_{st} \pm Z_{\alpha/2} \sqrt{V(\bar{y}_{st})}$$
$$3.9286 \pm 1.96 \sqrt{0.00008852} = (3.92, 3.96)$$

3. Find the minimum sample size n, such that estimate $\bar{y}_{st}$ is within the minimum bound ϵ with probability $1-\alpha$. Assume Neyman allocation is applied to each stratum. **(20 marks)**

$$P(|\bar{y}_{st} - \bar{Y}| < \epsilon) = 1-\alpha$$

$$\epsilon^2 = \text{var}(\bar{y}_{st}) Z^2_{\alpha/2}$$

$$\epsilon^2 = \sum_{h=1}^{H} W_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{S_h^2}{n_h} Z^2_{\alpha/2}$$

3 marks

$$\sum_{h=1}^{H} W_h^2 \left[\frac{1}{n_h} - \frac{1}{N_h}\right] S_h^2 = \frac{\epsilon^2}{Z^2_{\alpha/2}}$$

let $\frac{\epsilon^2}{Z^2_{\alpha/2}} = D$

5 marks

$$\therefore \sum \frac{W_h^2 S_h^2}{n_h} - \sum \frac{W_h^2 S_h^2}{N_h} = D$$

$$\sum \frac{W_h^2 S_h^2}{n_h} = D + \sum \frac{W_h^2 S_h^2}{N_h} \quad \cdots\cdots (1)$$

Under neyman allocation

substitute $n_h = \frac{n N_h S_h}{\sum N_h S_h}$ in (1) 2 marks

$$\therefore \frac{\sum W_h^2 S_h^2}{\left(\frac{n N_h S_h}{\sum N_h S_h}\right)} = D + \frac{\sum W_h^2 S_h^2}{N_h}$$

$$\sum \frac{N_h^2 S_h^2}{n N^2 N_h S_h} \left(\sum N_h S_h\right) = D + \frac{\sum W_h^2 S_h^2}{N_h}$$

10 marks

$$\frac{\left(\sum N_h S_h\right)\left(\sum N_h S_h\right)}{n N^2} = D + \frac{\sum W_h^2 S_h^2}{N_h}$$

$$\left(\sum N_h S_h\right)^2 = n \left[N^2 D + N^2 \sum \frac{N_h^2 S_h^2}{N^2 N_h}\right]$$

$$\left(\sum N_h S_h\right)^2 = n \left[N^2 D + \sum N_h S_h^2\right]$$

$$\therefore n = \frac{\left(\sum N_h S_h\right)^2}{N^2 D + \sum N_h S_h^2} = \frac{\left(\sum N_h S_h\right)^2}{N^2 \frac{\epsilon^2}{Z^2_{\alpha/2}} + \sum N_h S_h^2}$$

4. The advertising firm finds that obtaining an observation from rural household costs more than obtaining a response in town A or B. the increase is due to costs of traveling from one rural household to another. The cost per observation in each town is estimated to be \$9 (i.e. $c_1 = c_2 = 9$), and the costs per observation in the rural area to be \$16 ($i.e. c_3 = 16$). The stratum standard deviation (from a prior survey) are $\sigma_1 = 5$, $\sigma_2 = 15$, and $\sigma_3 = 10$. Find the overall sample size n and the stratum sample sizes n_1, n_2, and n_3, such that allow the firm to estimate the average TV-viewing time with $\epsilon = 2$ at minimum cost. Note that the population size in town A, B and rural area are respectively 155, 62, and 93.

(a) Assume using proportional allocation to each stratum **(5 marks)**

$$N = N_1 + N_2 + N_3 = 155 + 62 + 93 = 310$$

(2 marks) $$n = \frac{N(\sum N_h S_h^2)}{N\frac{\epsilon^2}{Z_{\alpha/2}^2} + \sum N_h S_h^2} = \frac{310 \times (155 \times 5^2 + 62 \times 15^2 + 93 \times 10^2)}{310 \times \frac{2^2}{1.96^2} + (155 \times 5^2 + 62 \times 15^2 + 93 \times 10^2)} = 66$$

$$n_1 = \frac{N_1}{N} \times n = \frac{155}{310} \times 66 = 33 \quad (1 \text{ mark})$$

$$n_2 = \frac{N_2}{N} \times n = \frac{62}{310} \times 66 = 13.2 = 13 \quad (1 \text{ mark})$$

$$n_3 = \frac{N_3}{N} \times n = \frac{93}{310} \times 66 = 19.8 \simeq 20 \quad (1 \text{ mark})$$

(b) Assume using optimum allocation to each stratum **(5 marks)**

$$n = \frac{(\sum N_h S_h / \sqrt{C_h})(\sum N_h S_h \sqrt{C_h})}{N^2 \frac{\epsilon^2}{Z_{\alpha/2}^2} + \sum N_h S_h^2}$$

(2 marks) $$= \frac{\left(\frac{155 \times 5}{\sqrt{9}} + \frac{62 \times 15}{\sqrt{9}} + \frac{93 \times 10}{\sqrt{16}}\right)\left(155 \times 5 \times \sqrt{9} + 62 \times 15 \times \sqrt{9} + 93 \times 10 \times \sqrt{16}\right)}{310^2 \times \frac{2^2}{1.96^2} + (155 \times 5^2 + 62 \times 15^2 + 93 \times 10^2)}$$

$$\therefore n = 55.63 \simeq 56$$

$$n_1 = n\left(\frac{N_1 S_1 / \sqrt{C_1}}{\sum_{h=1} N_h S_h / \sqrt{C_h}}\right) = 17.9 = 18 \quad (1 \text{ mark})$$

$$n_2 = 21.5 = 22 \quad (1 \text{ mark})$$

$$n_3 = 16.15 = 16 \quad (1 \text{ mark})$$

(c) Assume using neyman allocation to each stratum **(5 marks)**

$$n = \frac{(\sum N_h S_h)^2}{N^2 \frac{\epsilon^2}{Z_{\alpha/2}^2} + \sum N_h S_h^2} = \frac{(155 \times 5 + 62 \times 15 + 93 \times 10)^2}{310^2 \times \frac{2^2}{1.96^2} + (155 \times 5^2 + 62 \times 15^2 + 93 \times 10^2)}$$

$$= 54.589 \simeq 55 \quad (2 \text{ marks})$$

$$n_h = \frac{n N_h S_h}{\sum N_h S_h}$$

$n_1 = 16$ (1 mark)

$n_2 = 19$ (1 mark)

$n_3 = 19$ (1 mark)