Assignment 9

A9 – 1. A sample of 692 families was selected from a large region to estimate, among other things, $\mathbf{P}$, the proportion of families with vegetable gardens in a particular year. The families were classified into three strata – urban, rural non-farm, and farm – because these groups were expected to show differences in the frequency and size of vegetable gardens. The sampling fraction was roughly the same in all strata, a sample of 1 per 1,000 being aimed at. The values of the population weights ($\mathbf{W}_h$), the stratum sample sizes (n_h), and numbers of families with vegetable gardens m_h) were as shown in the table at the upper right.

Stratum (h)	$\mathbf{W}_h$	n_h	m_h
Urban	0.445	300	218
Rural non-farm	0.230	155	147
Farm	0.325	237	229
Total	**1.000**	**692**	**594**

(a) Estimate $\mathbf{P}$ and find an approximate 95% confidence interval for this proportion.

(b) If the costs per response are the same in all strata, find the approximate optimum allocation for a total sample size of 692 families for such a sample survey, and an approximate value for the standard deviation of the estimate of the population proportion that it would yield.

(c) Ignoring the stratification, suppose that the data had come from observing 594 families with vegetable gardens among 692 families obtained by equiprobable selecting. Estimate $\mathbf{P}$ and find an approximate 95% confidence interval for this proportion. Comment briefly on the precisions of the estimates in (a), (b) and (c).

(d) Suppose that the costs per response in the strata are \$4, \$9 and \$10 respectively. If a total budget of \$4,000 is vailable and the fixed costs are \$800, find the total sample size and the approximate optimum allocation.

A9 – 2. An opinion poll asks whether or not individuals are in favour of a sales tax on grocery purchases. The large study population is stratified by sex (Male, Female) and by income (Low, Medium or high); the population stratum weights are as shown in the *upper* table at the right. A pilot survey of 20 people obtained by equiprobable selecting from *each* of the four strata yields the **proportions in favour** of the tax shown in the *lower* table at the right.

	INCOME	
SEX	Low	Medium or high
Male	0.1	0.4
Female	0.2	0.3

	INCOME	
SEX	Low	Medium or high
Male	0.10	0.20
Female	0.05	0.15

(a) The whole sample survey will involve choosing a total of 500 people; *i.e.*, the 80 from the pilot survey plus 420 more. Find the approximate optimal allocation of these 500 among the strata for estimating the population proportion of people in favour of the tax.

(b) What assumption(s) are made in (a)? Explain briefly whether they are reasonable.

A9 – 3. An opinion poll asks whether or not individuals are in favour of a sales tax on grocery purchases. The large study population is stratified by sex (Male, Female) and by income (Low, Medium, High); the population stratum weights are as shown in the *upper* table at the right. One hundred people obtained by equiprobable selecting from *each* stratum yield the **proportions in favour** of the tax given in the *lower* table at the right.

		INCOME	
SEX	Low	Medium	High
Male	0.08	0.33	0.05
Female	0.12	0.35	0.07

		INCOME	
SEX	Low	Medium	High
Male	0.10	0.10	0.25
Female	0.03	0.12	0.20

(a) Estimate the proportion of people in the study population in favour of the tax and find an approximate 99% confidence interval for this proportion.

(b) Was the stratification effective? Justify your answer briefly.

A9 – 4. Dairy farms in a certain region were divided into four categories, depending on their total acreage and on whether or not they concentrated exclusively on dairy products; the respective numbers of farms in the four categories were 72, 37, 50 and 21. In a sample survey to estimate the number of milk-producing cows in the region, and the proportion of them which were diseased, 30 farms were obtained by equiprobable selecting from the four categories taken as strata with (roughly) proportional allocation. The data obtained were as follows:

Category (h)		Numbers of cows on farms chosen (y_{hj})	Σy_{hj}	Σy_{hj}^2	n_h
1	Total number	61 47 44 70 28 39 51 52 101 49 54 71	667	40,935	12
	Number diseased	2 1 3 2 0 1 2 1 2 0 4 2	20	48	
2	Total number	160 148 89 139 142 93	771	103,559	6
	Number diseased	3 4 2 1 6 2	18	70	
3	Total number	26 19 21 34 28 15 20 24	187	4,619	8
	Number diseased	1 0 1 2 1 0 1 2	8	12	
4	Total number	17 11 12 15	55	779	4
	Number diseased	0 1 1 0	2	2	

(*continued overleaf*)

A9 – 4. (cont.) (a) Estimate the total number of cows on the 180 farms in the region, and give an appropriate measure of the precision of your estimate.

*(b) Estimate the proportion of diseased cows on the 180 farms in the region, and give an appropriate measure of the precision of your estimate.

(c) If you were asked for a Plan for a future sample survey of the number of cattle in the region with the same sample size, explain briefly the change(s) you would make to the sample allocation if the cost per response is the same in each category; give the category (stratum) sample sizes for the allocation you would recommend.

***A9 – 5.** A large company is organized in 11 departments *A*, *B*, ..., *K*; seven of these departments are spread over two locations; the number of employees in each department is shown by location in the second lines of the two tables below. An employee opinion sample survey uses the two locations as strata and equiprobable selecting within the strata of departments as clusters. One response is employee opinion as to how well management communicates to employees its objectives for the company; this response is measured for all employees in the eight departments selected by combining scores to three questions with answers on a 5-point Likert scale. The average and standard deviation of this response are given by department in the third and fourth lines of the two tables.

LOCATION 1:

Department	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*
Number of employees	230	110	25	322	17	65	63	210	77	12	45
Sample average	3.62				4.24	4.07		3.36	3.81		
Sample standard deviation	0.87				1.03	0.75		1.44	1.31		

LOCATION 2:

Department	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Number of employees	410	50	70	225	110	37	63
Sample average		4.48	4.15	4.03			
Sample standard deviation		0.58	0.77	0.83			

(a) Prepare a scatter diagram of the sample data for the two locations; also plot sample average score against number of employees in the department. Identify, and suggest an explanation for, the important feature(s) of the diagram. [The data for Location 1 are those in question **A6 – 4**; again assume there is complete response by department.]

(b) Find an approximate 90% confidence interval for the *difference* in average score between the two locations.

(c) Find an approximate 90% confidence interval for the average score across the whole company.

A9 – 6. A stratified population of total size $\mathbf{N}$ is made up of k strata of sizes $\mathbf{N}_1$, $\mathbf{N}_2$, $\mathbf{N}_3$, ..., $\mathbf{N}_k$; stratum h (h = 1, 2, ..., k) contains a proportion $\mathbf{P}_h$ of units possessing a particular characteristic *C*.

(a) If $n \ll \mathbf{N}$, show that, when the stratum sample sizes are *optimally* allocated under equiprobable selecting from the stratified population, the square of the standard deviation of the estimator of the population proportion ($\mathbf{P}$) of units with characteristic *C* is approximately as shown at the right.

$$\frac{1}{n}\Big[\sum_{h=1}^{k}\frac{\mathbf{N}_h}{\mathbf{N}}\sqrt{\mathbf{P}_h(1-\mathbf{P}_h)}\Big]^2$$

(b) Find the corresponding expression to that in (a) for *proportional* allocation and for *equal* allocation.

***A9 – 7.** Consider the case k = 2 of the previous question **A9 – 6**.

(a) Give expressions for the efficiency of the estimator of $\mathbf{P}$ under equiprobable selecting from the stratified population for proportional allocation and for equal allocation, compared with optimum allocation of stratum sample sizes.

(b) Regarding the $\mathbf{P}_h$ as fixed, find the *minimum* efficiency of proportional and of equal allocation, compared with optimum allocation of stratum sample sizes. Illustrate how these minima change with different values of the $\mathbf{P}_h$ for a representative range of values of the $\mathbf{P}_h$.

(c) How is the solution to (b) affected if the $\mathbf{N}_h$, rather than the $\mathbf{P}_h$, are regarded as fixed?

*Optional question for enrichment