Assignment 6

A6 – 1. A manufacturer of band saws wants to estimate the average repair cost per month for the saws he has sold to certain industries. He cannot obtain a repair cost per saw, but he has access to the total amount spent for saw repairs and the number of saws owned by each industry; he therefore decides to use each industry as a *cluster.* The manufacturer obtains, by equiprobable selecting, 20 of the 96 industries he services. The data were as follows, where the 'Repair cost' is the total repair cost for the past month in dollars:

Industry	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Number of saws	3	7	11	9	2	12	14	3	5	9	8	6	3	2	1	4	12	6	5	8
Repair cost	50	110	230	140	60	280	240	45	60	230	140	130	70	50	10	60	280	150	110	120 .

On the basis of these data, find an approximate 90% confidence interval for:

(a) the average repair cost per saw for the past month;

(b) the total amount spent by the 96 industries on band saw repairs during the past month.

(c) Repeat (b) if the manufacturer later finds from his records that he has sold a total of 710 saws to the 96 industries.

A6 – 2. An economic survey is designed to estimate the average amount spent per month on utilities by households in a city; because no proper list of households is available, it is decided to use city blocks as *clusters.* A map of the city is divided into 60 blocks and 20 of these are obtained by equiprobable selecting. Interviewers then obtain the cost of utilities from *all* households in each of the chosen blocks, with the following results:

Block (j)	Number of households (l_j)	Total amount spent (${}_Ty_j$) on utilities	Block (j)	Number of households (l_j)	Total amount spent (${}_Ty_j$) on utilities
1	55	$4,420	11	73	$5,860
2	60	4,780	12	64	4,940
3	63	4,860	13	69	5,660
4	58	4,760	14	58	4,740
5	71	5,520	15	63	4,780
6	78	6,220	16	75	5,740
7	69	5,560	17	78	6,420
8	58	4,740	18	51	4,860
9	52	3,980	19	67	5,460
10	71	5,620	20	70	5,760 ;

for these data:

$\sum_{j=1}^{m} l_j = 1{,}303$, $\sum_{j=1}^{m} l_j^2 = 86{,}171$;

$\sum_{j=1}^{m} {}_Ty_j = 104{,}680$, $\sum_{j=1}^{m} {}_Ty_j^2 = 555{,}494{,}400$;

$\sum_{j=1}^{m} {}_Ty_j l_j = 6{,}912{,}460$.

(a) Estimate the *average* amount per month a household in the city spends on utilities, and give an approximate 99% confidence interval for this average.

(b) Estimate the *total* amount per month spent on utilities by households in the city, and give an approximate 99% confidence interval for this total.

A6 – 3. A car rental company has offices (with accompanying car depots) in 230 cities. To study certain operating characteristics of its fleet of cars, the company obtains 20 of the offices by equiprobable selecting and, from complete office records, compiles the following data on the number of cars in each depot, the total mileage driven by these cars in the last year, and the number of cars needing major repairs:

Number of cars	Total mileage	Number repaired	Number of cars	Total mileage	Number repaired
6	145,920	1	12	342,960	2
2	54,120	0	6	163,620	1
11	303,600	1	5	102,300	1
7	196,560	1	3	55,140	1
8	220,480	1	7	160,580	1
14	406,980	2	7	130,270	1
6	192,180	1	2	44,100	0
2	56,820	1	4	62,560	1
2	57,820	0	2	40,860	1
5	127,750	1	8	156,400	2 .

The numerical summaries for these data are:

Number of cars: $\sum_{j=1}^{m} l_j = 119$,

$\sum_{j=1}^{m} l_j^2 = 939$;

Total mileage: $\sum_{j=1}^{m} {}_Ty_j = 3{,}021.020$,

$\sum_{j=1}^{m} {}_Ty_j^2 = 661{,}743{,}493{,}000$;

Number repaired: $\sum_{j=1}^{m} {}_Ty_j = 20$,

$\sum_{j=1}^{m} {}_Ty_j^2 = 26$;

Cross products: $\sum_{j=1}^{m} {}_Ty_j l_j = 24{,}663{,}420$;

$\sum_{j=1}^{m} {}_Ty_j l_j = 147$.

(a) Find an approximate 95% confidence interval for the *average* annual mileage driven per car for the company's fleet.

(b) Find an approximate 95% confidence interval for the *total* annual mileage driven by the company's fleet.

(c) Estimate the proportion of the company's cars which needed major repairs during the year, and find an approximate 95% confidence interval for it.

(*continued overleaf*)

A6 – 4. A large company is organized in 11 departments *A*, *B*, ..., *K*; the number of employees in each department is shown in the second line of the table below. As part of an employee opinion sample survey, data are complied for entire departments obtained from the 11 by equiprobable selecting. One response of interest 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 five departments selected by combining scores to three questions with answers on a 5-point scale. Data summaries for this response for each of these departments are given in the third and fourth lines of the following table:

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		.

(a) Prepare a scatter diagram of the sample data for the five departements; plot sample average score against number of employees in the department. Identify, and suggest an explanation for, the important feature(s) of the diagram.

(b) Use the estimator $\overline{Y}_{uc}$ to find an approximate 99% confidence interval for the average score of *all* the company's employees.

(c) When selecting *un*equal sized clusters, ${}_{\mathrm{T}}Y = \mathbf{M}_{\mathrm{T}}\overline{Y}$ can be used to estimate a respondent population *total* when $\mathbf{N}$ is *un*known; here, $\mathbf{N}$ is *known* so use ${}_{\mathrm{T}}Y/\mathbf{N}$ to find (again) an approximate 99% confidence interval for the *average* score of all the company's employees.

(d) Explain briefly how the confidence intervals obtained in (b) and (c) illustrate more or less effective management of variation in the Analysis stage of the FDEAC cycle.

A6 – 5. A large corporation is considering revising its retirement policy, and wishes to estimate the proportion and the number of its employees in favour of the new policy. The corporation consists of 87 separate plants spread across the country; because results must be obtained as soon as possible and at minimal cost, it is decided to regard each plant as a *cluster*. Fifteen plants are obtained by equiprobable selecting (EPS) and the opinions of all the employees at each plant are gathered, with the following results:

Plant number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of employees	51	62	49	73	101	48	65	49	73	61	58	52	65	49	55
Number in favour	42	53	40	45	63	31	38	30	54	45	51	29	46	37	42 ;

numerical summaries of these data are:

Total number of employees: 911 [with a corresponding sum of squares of 58,075],
Total number in favour: 646 [with a corresponding sum of squares of 29,104],
The sum of products is: 40,730.

(a) Find an approximate 99% confidence interval for the *proportion* of the corporation's employees who are in favour of the new policy.

(b) Find an approximate 99% confidence interval for the *number* of the corporation's employees who are in favour of the new policy.

(c) Explain briefly whether you would expect the estimate in (a) to be more or less precise than an estimate based on obtaining *individually*, by EPS, the same number of employees. Check your answer by an appropriate calculation based on the data given above.

A6 – 6. (a) Compare and contrast, from the perspective of the Analysis stage of the FDEAC cycle, the *mathematical* structure in Question **A6 – 5** above and that in Question **A4 – 2** about taxi tires.

(b) Compare and contrast, from the perspective of the Analysis stage of the FDEAC cycle, the *mathematical* structure in Question **A6 – 5** above with that in Question **A4 – 3** about insect larvae in a stream bed.

(c) In light of your answers to (a) and (b) and also to Question **A4 – 7**, compare and contrast these *mathematical* structures on the basis of our discussion of clustered sample survey designs and estimating a ratio.

A6 – 7. To estimate the proportion of diseased plants in a large field of corn, a farmer obtains, by equiprobable selecting (EPS), 40 clusters of nine plants; each cluster is made up of 3 plants in each of 3 neighbouring rows. The numbers of diseased plants (out of 9) in the 40 clusters are found to be as follows:

2 5 1 1 1   7 0 0 3 2   3 0 0 0 7   0 4 1 2 6   0 0 1 4 5   0 1 4 2 6   0 2 4 1 7   3 5 0 3 6.

Find an approximate 95% confidence interval for the proportion of diseased corn plants in the field:

(a) (i) on the basis of selecting *clusters*, as actually done to gather the data;
(ii) ignoring the clustering and assuming that the data are from 360 corn plants obtained *individually* by EPS.

(b) Explain briefly how the relative precision of EPS of *clusters* and EPS of *individual* corn plants is related to the type of distribution of the disease throughout the field.

(c) Give, in point form, a set of instructions a farmer could follow to implement the clustered design in this situation.