Assignment 7

A7 – 1. A forest resource manager is interested in estimating the number of dead fir trees in a 300-acre area of heavy parasite infestation. Using an aerial photo, she divides the area into 200 one-and-one-half acre plots and, from an examination of the photo, estimates that there are a total of 4,200 dead fir trees in the area. From a ground count in 10 of the 200 plots obtained by equiprobable selecting, the actual number of dead fir trees, together with the corresponding estimates from the aerial photo, are as shown above at the right. Use these data to estimate the total number of dead fir trees in the 300 acres, and give an appropriate measure of the precision of your estimate.

Plot sampled	1	2	3	4	5	6	7	8	9	10
Photo count	12	30	24	24	18	30	12	6	36	42
Ground count	18	42	24	36	24	36	14	10	48	54.

A7 – 2. In an investigation study of lung function in a group of 560 coal miners, an estimate was required of the average of a particular index ($\mathbf{Y}$ of lung function. Ten miners were obtained by equiprobable selecting and their $\mathbf{Y}$-values, y_j, were measured; a note was also made of each miner's height, h_j. The data obtained are shown above at the right. From medical records, the average height for the group of 560 miners was known to be 173.2 cm. Using these data, estimate the average of $\mathbf{Y}$ and give an appropriate measure of the precision of your estimate; use both ratio estimating and univariate estimating ignoring the auxiliary variable, and comment briefly on similarities or differences in the resulting estimates and their precision.

Miner (j)	1	2	3	4	5	6	7	8	9	10
Lung function (y_j)	3.0	3.5	3.3	3.1	4.1	3.2	3.7	2.9	3.9	3.4
Height (h_j cm)	173	183	170	175	160	157	168	180	178	163.

A7 – 3. Part of a coniferous forest contains 280 trees of the same species and of similar ages. A preliminary estimate is required of the total weight of lumber that these trees will yield. A forestry expert claims to be able to make fairly accurate assessments of the yield from any tree merely by visual inspection, and he makes such assessments for all 280 trees. His assessment is a total yield of 432.6 tonnes of lumber. Subsequently, 25 trees obtained by equiprobable selecting are felled and their lumber yields accurately determined. These actual yields (y_j), and the corresponding assessed yeilds (x_j), provide the numerical summaries given at the right, where the units are tonnes or tonnes^2

$$\sum_{j=1}^{n} x_j = 40.7 \quad , \quad \sum_{j=1}^{n} x_j^2 = 71.17;$$
$$\sum_{j=1}^{n} y_j = 39.2 \quad , \quad \sum_{j=1}^{n} y_j^2 = 66.92;$$
$$\sum_{j=1}^{n} y_j x_j = 68.43 \;, \quad n = 25.$$

(a) Estimate the total yield of lumber from the 280 trees, using ratio estimating, and find an approximate 95% confidence interval for this total.

(b) Estimate the total yield of lumber from the 280 trees, based *only* on the sample y_j-values, and gain find an approximate 95% confidence interval or this total.

(c) Discuss briefly which of the two estimates in (a) and (b) you consider to have fewer limitations.

A7 - 4. A sum of \$8,000 is available to conduct a sample survey of weekly expenditure on food by households in a certain town. A list of households is available; there are 8,220 in all. The investigation is to relate to a specific week of the year. Interview cost to determine the week's expenditure (y_j) for a household selected equiprobably \$6.50; it costs an additional \$0.50 to also obtain the value yields (x_j) of a relevant auxiliary variate. The cost of setting up the survey is \$1,000. The population average, $\overline{\mathbf{X}}$, is known to be 29. Previous experience suggests that other population attributes will have approximately the values given at the right. To estimate the respondent population average, $\overline{\mathbf{Y}}$, on the basis of equiprobable selecting, determine whether it would be better to use a ratio estimator tor a univariate estimator without reference to the auxiliary variate, if the respective data processing costs are \$200 and \$100 for the two calculations. What sample size should be used? Set out your reasoning and calculations clearly.

$$\mathbf{S}_{\mathbf{X}}/\overline{\mathbf{X}} \simeq 0.2$$
$$\mathbf{S}_{\mathbf{Y}}/\overline{\mathbf{Y}} \simeq 0.3$$
$$\mathbf{S}_{\mathbf{YX}}/\mathbf{S}_{\mathbf{Y}}\mathbf{S}_{\mathbf{X}} \simeq 0.45$$
$$\mathbf{R} = 3.1$$

A7 – 5. From a large population of families, 33 are obtained by equiprobable selecting (EPS), from and complete response is obtained for both the family size ($\mathbf{X}$) and the dollar expenditure per week on food ($\mathbf{Y}$); the sample data yield the numerical summaries given at the right. Find an approximate 99% confidence interval for the average expenditure on food per person per week in the population of families.

$$\sum_{j=1}^{n} x_j = 123 \quad , \quad \sum_{j=1}^{n} x_j^2 = 533;$$
$$\sum_{j=1}^{n} y_j = 2{,}721.30 \quad , \quad \sum_{j=1}^{n} y_j^2 = 254{,}196.00;$$
$$\sum_{j=1}^{n} y_j x_j = 10{,}786.50 \;, \quad n = 33.$$

A7 – 6. A study population consists of $\mathbf{N} = 5$ elements which have the values of two responses ($\mathbf{Y}$ and $\mathbf{X}$) as in the table at the right.

(a) List all the possible samples of size 2 that can be obtained from this population under EPS.

(b) For each sample, tabulate the value of both $\bar{y}_{un}$ (the sample average, which is the 'univariate' estimate of $\overline{\mathbf{Y}}$, the respondent population average) and $\bar{y}_R$ (the 'bivariate' estimate of $\overline{\mathbf{Y}}$); also, tabulate the probability of selecting each sample.

i	$\mathbf{Y}$	$\mathbf{X}$
1	1	3
2	3	5
3	3	4
4	2	4
5	5	6

(*continued overleaf*)

A7 – 6. (cont.) (c) Use the relevant probability function from (b) to find the *estimating bias* of the random variable $\overline{Y}_R$ under EPS; then calculate an approximate value using equation (2.17.12) from page 3.131 in Figure 2.17 given at the right.

$$E(\overline{Y}_R) - \overline{\mathbf{Y}} \simeq \frac{1}{\overline{\mathbf{X}}}(\mathbf{R}\mathbf{S}_{\mathbf{X}}^2 - \mathbf{S}_{\mathbf{YX}})(\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})$$

$$e.b.(R) \simeq \frac{1}{\overline{\mathbf{X}}^2}\big[\mathbf{R}\mathbf{S}_{\mathbf{X}}^2 - \mathbf{S}_{\mathbf{YX}}\big](\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})$$

(i) Discuss briefly how closely the two values of this estimating bias agree.

(ii) Show how the upper expression at the right for the estimating bias of $\overline{Y}_R$ under EPS is derived from (below it) equation (2.15.7) near the middle of the second side (page 2.114) of Figure 2.15 of the Course Materials.

(d) Use the relevant probability function in (b) to find, under EPS, the square of the standard deviation of the random variable representing the ratio estimator, as defined at the upper right; then use equation (2.17.6) on page 2.131 in Figure 2.17 (given at the lower right) to calculate an approximate value of this standard deviation. Discuss briefly how closely the two values of this (squared) s.d. agree.

$$[s.d.(\overline{Y}_R)]^2 = E[\overline{Y}_R - E(\overline{Y}_R)]^2$$

$$[s.d.(\overline{Y}_R)]^2 \simeq \big[\mathbf{S}_{\mathbf{Y}}^2 - 2\mathbf{R}\mathbf{S}_{\mathbf{YX}} + \mathbf{R}^2\mathbf{S}_{\mathbf{X}}^2\big](\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})$$

(e) Use the probability functions in (b) to find, under EPS, the square of the standard deviation of the random variable representing the sample average, as defined at the right; then, using the result in (d), find the difference $[s.d.(\overline{Y}_{un})]^2 - [s.d.(\overline{Y}_R)]^2$. Finally, use (at the right below) equation (2.17.18) on page 2.133 in Figure 2.17 to calculate an approximate value of this difference.

$$[s.d.(\overline{Y}_{un})]^2 = E(\overline{Y}_{un} - \overline{\mathbf{Y}})^2$$

(i) Discuss briefly how closely the two values of the difference agree.

$$[s.d.(\overline{Y}_{un})]^2 - [s.d.(\overline{Y}_{bi})]^2 \simeq \mathbf{R}\big[2\mathbf{S}_{\mathbf{YX}} - \mathbf{R}\mathbf{S}_{\mathbf{X}}^2\big](\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})$$

(ii) Discuss briefly the efficiency of ratio estimating compared with univariate estimating of $\overline{\mathbf{Y}}$ in this situation.

***A7 – 7.** To estimate the total wheat yield of a province by means of a sample survey, the province is divided into four regions, with individual farms as the sample units in each region. The number of farms in each region is given in the sixth column of the table above. An aerial survey of the whole province is then made to estimate the acreage under wheat; from these acreage estimates, the aerial estimate of the total wheat yield of the province is 20×10^6 bushels. Further, two farms are obtained by EPS from each region; for these farms, aerial estimates of wheat yields as well as ground estimates (from on-site inspection) are obtained (in bushels) as given in the table above.

REGION NUMBER	AERIAL ESTIMATES Farm 1	Farm 2	GROUND ESTIMATES Farm 1	Farm 2	NO. OF FARMS IN REGION
1	2,100	5,200	1,870	4,920	2,000
2	1,600	2,300	1,450	2,000	1,400
3	2,400	1,200	2,060	1,120	4,000
4	1,900	3,100	1,710	2,730	600 .

(a) Estimate the total wheat yield of the province and its standard deviation. Comment briefly on how your method of calculation would be expected to affect the accuracy of your estimate, and explain whether its standard deviation should be used as the basis of an approximate confidence interval for the total wheat yield of the province.

(b) Using the ground estimates only and the estimated stratum squared s.d.s given at the right, find the Neyman allocation for a total sample size of 200 farms from the province.

Region number	1	2	3	4
Estimated s.d.^{2} ($\times10^6$)	1.0	0.2	0.4	0.5.

Also estimate the standard deviation of an estimate of the total provincial wheat yield for this sample survey Plan.

***A7 – 8.** A company that sells television sets has 123 district offices; we define the response $\mathbf{Y}_i$ to be the earnings in office i in the first quarter of *this* year and $\mathbf{X}_i$ to be the earnings in office i in the first quarter of *last* year. It is known that the total earnings of the company in the first quarter of last year were ${}_T\mathbf{X} = 128{,}200$ in suitable units (so that $\overline{\mathbf{X}} = 128{,}200/123 \simeq 1{,}042.2764$ units). For n = 13 of the company's offices obtained by equiprobable selecting, the sales data (in the same units) for the two periods were:

j	1	2	3	4	5	6	7	8	9	10	11	12	13
y_j	610	780	1,600	1,030	600	1,050	977	1,440	1,570	2,210	980	865	1,710
x_j	550	720	1,500	1,020	620	980	928	1,200	1,350	1,750	670	729	1,530;

for these data: $\sum_{j=1}^{n} y_j = 15{,}422$, $\sum_{j=1}^{n} y_j^2 = 21{,}073{,}754$; $\sum_{j=1}^{n} x_j = 13{,}547$, $\sum_{j=1}^{n} x_j^2 = 15{,}963{,}525$; $\sum_{j=1}^{n} y_j x_j = 18{,}273.841$, n = 13.

Estimate the aversge sales per office for the company and give an approximate 99% confidence interval for this average, assuming that the selection of the 13 offices was based on:

(a) equiprobable selecting (EPS) from from the *un*stratified population and the data for $\mathbf{Y}$ only.

(b) EPS from the stratified population and using the data for $\mathbf{Y}$ only only – use two strata, with current sales of less than and more than 1,000 units, and take the respective stratum sizes as 57 and 66 district offices;

(c) EPS from the *un*stratified population and using a ratio estimator – your solution should include a scatter diagram of the data and the coefficient of variation of $\overline{X}$, together with appropriate comments on them, and also a check of the relative precision of univariate and bivariate estimating in this situation.

(d) the same stratification as in (b) but using ratio estimators, both combined and separate – explain briefly what additional information is desirable in this case;

(e) EPS from the *un*stratified population and using a linear regression estimator.

In light of differences among the five estimators, explain briefly which is the estimator of choice in this situation.

*Optional question for enrichment