FE 1

Last name, first name: Student #:

**UNIVERSITY OF TORONTO**
**Faculty of Arts and Science**

**AUGUST EXAMINATIONS 1997**

**STA 322S**

**Duration - 3 hours**

**Examination Aids: Non-Programmable Calculator, aid sheet, both sides, with theoretical formulas only.**

[20] 1) A sociological study conducted in a small town is interested in the age structure of the residents. The city has 620 households. An SRS of n = 10 households was selected from the directory. At the completion of the fieldwork the following results were obtained:

| household | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size | 6 | 4 | 5 | 7 | 3 | 2 | 8 | 4 | 4 | 5 | 48 |
| children (under 18) | 3 | 2 | 3 | 3 | 1 | 0 | 4 | 2 | 0 | 2 | 20 |
| adults over 65 | 2 | 0 | 0 | 2 | 0 | 2 | 1 | 0 | 2 | 2 | 11 |

(a) Estimate the following parameters, and place a bound on the error of estimation:
- (1) the total number of residents in the town
- (2) the total number of adults (over 18) in the town
- (3) the total number of households that contain at least one resident over 65.

(b) How large a sample should be taken in order to estimate the proportion of households that contain at least one resident over 65 with a bound of 0.1 on the error of estimation?

6 (a) (1) $\hat{\tau} = N\bar{y} = 620 \times \frac{48}{10} = \underline{2976}$ 1, $\bar{y} = 4.8$, $s^2 = 3.289$

$\hat{SD}(\hat{\tau}) = N\,\hat{SD}(\bar{y}) = 620 \times \sqrt{\frac{620-10}{620}\frac{s^2}{10}} = \underline{352.7}$, $B = 2 \times 352.7$ 1 $= \underline{705.4}$

(2) adults (over 18) = size - children

$\bar{y} = \frac{28}{10} = 2.8$, $\hat{\tau} = N\bar{y} = 620 \times 2.8 = \underline{1736}$ 1

$s^2 = 0.844$, $\hat{SD}(\hat{\tau}) = \underline{178.7}$ 1 $B = 2 \times 178.7 = 357.4$

(3) $p$ — prop. of households that contain at least one resident over 65

$\hat{p} = \frac{6}{10}$ , $\hat{\tau} = N\hat{p} = 620 \times 0.6 = 372$ 1

$\hat{SD}(\hat{\tau}) = N\hat{SD}(\hat{p}) = 620\sqrt{\frac{\hat{p}\hat{q}}{n-1}\frac{N-n}{N}} = 620\sqrt{\frac{0.6 \times 0.4}{10-1}\frac{610}{620}}$

$= 620 \times 0.16197 = 100.4$ 1

$B = 200.8$ (even if $p = 0.5$, is used it can be accepted)

2 (4) wrong presample result $\hat{p} = 0.6$

$B = 0.1$ , $D = \left(\frac{B}{2}\right)^2 = \left(\frac{0.1}{2}\right)^2$ ,

$n = \frac{Npq}{(N-1)D + pq} = \frac{620 \times 0.24}{619 \times \left(\frac{0.1}{2}\right)^2 + 0.24} = 83.2$

FE 2

(c) Estimate the proportion of children under 18 in the town and place a bound on the error of estimation. Is this estimator unbiased? Explain.

(d) What do you need to calculate an unbiased estimator in (c)? Would you prefer to use that unbiased estimator instead of one already used in (c)? Explain.

(e) Assuming that the total # of residents is 3100, propose the best estimator you can apply to the sample to estimate the average # of children under 18 per household, calculate it and place a bound on the error of estimation.

4 (c) $r = \hat{p} = \frac{\sum y_i}{\sum x_i} = \frac{20}{48} = 0.417 = 41.7\%$ 2

- ratio estimator, biased

pop. size is not known, so that

$\hat{V}(\hat{p}) = \frac{N-n}{N} \frac{s_r^2}{\bar{x}^2 n} = \frac{610}{620} \frac{0.5895}{(4.8)^2 10} = 2.5173 \times 10^{-3}$

$\hat{SD}(\hat{P}) = 0.050$ , $B = 0.10$ 2

$s_r^2 = \frac{1}{n-1} \sum (y_i - r x_i)^2 = \frac{1}{9} [(3-0.417\times 6)^2 + \cdots + (2-0.417\times 5)^2 = 0.5895$

3 (d) We need the number of residents in the town, and then $\hat{p} = \frac{\hat{\tau}}{\tau_x} = \frac{N\bar{y}}{\tau_x}$.

x and y are correlated, so that the ratio estimator in (c) is preferable.



FE 3

5 (e) The best estimator is regression estimator

$$\hat{\mu}_{yL} = \bar{y} - b(\bar{x} - \mu_x) \quad , \mu_x = \frac{3100}{620} = 5$$

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} = \frac{115 - 10 \times 4.8 \times 2}{260 - 10 \times (4.8)^2} =$$

$\bar{y} = 2, \ \bar{x} = 4.8$ $\qquad = \dfrac{19}{29.6} = 0.642$

$\Sigma x_i y_i = 115$

$$\hat{\mu}_{yL} = 2 - 0.642 \times (4.8 - 5) = \underline{\underline{2.13}} \quad 3$$

$s_x^2 = 3.289, \quad s_y^2 = 1.778$ $\qquad$ ($\frac{n-1}{n-2}$ can be ignored)

$$\hat{V}ar(\hat{\mu}_{yL}) = \frac{N-n}{N} \frac{1}{n} \frac{n-1}{n-2} [s_y^2 - b^2 s_x^2] =$$

$$= \frac{610}{620} \frac{1}{10} \frac{9}{8} [\underbrace{1.778 - 0.642^2 \times 3.289}_{0.4224}] = 0.04675$$

$\hat{SD}(\hat{\mu}_{yL}) = \underline{\underline{0.216}}$ , $\underline{\underline{B = 0.432}}$ $\quad 2$

FE4

[20] 2) Results for the number of man-hour lost, for a given month among laborers, technicians and administrators in a certain company are given in the following table:

| | Laborers | Technicians | Administrators |
|---|---|---|---|
| $N_i$ | 240 | 130 | 50 |
| $\Sigma x_i$ | 3360 | 1040 | 250 |
| $\Sigma x_i^2$ | 55680 | 11570 | 2050 |

(a) Consider laborers, technicians and administrators as a stratification of the population of employees. Calculate the strata means $\mu_i$, $\sigma_i^2$ and the population mean and variance $\mu$ and $\sigma^2$.

(b) Do you expect any significant advantage of using stratified sampling from this population than from using SRS? Explain. Do you expect any significant advantage of using proportional allocation instead of optimal allocation of the sample? Explain .(use the results from (a))

4 (a)

| | | L | T | A | comp |
|---|---|---|---|---|---|
| | $N_i$ | 240 | 130 | 50 | $420 = N$ |
| 2 | $\mu_i$ | 14 | 8 | 5 | $11.071 = \mu$ |
| 2 | $\sigma_i^2$ | 36 | 25 | 16 | $42.433 = \sigma^2$ |

$$\mu_i = \frac{\Sigma x_i}{N_i}, \quad \sigma_i^2 = \frac{1}{N_i}\Sigma x_i^2 - \mu_i^2, \quad \mu = \frac{\Sigma\Sigma x_i}{N}$$

$$\sigma^2 = \frac{1}{N}\Sigma\Sigma x_i^2 - \mu^2 \quad \left(\text{or } \sigma^2 = \Sigma\frac{N_i}{N}(\mu_i - \mu)^2 + \Sigma\frac{N_i}{N}\sigma_i^2, \quad \mu = \Sigma\frac{N_i}{N}\mu_i\right)$$

4(b) Differences between strata are not small (in means 14, 8, 5) so that stratified sample will be better than SRS    2    FES

Differences between strata variances $b_1 = 6$, $b_2 = 5$, $b_3 = 4$ 2 are not large. Optimal allocation will not give much better result than proportional allocation.
(someone would think that differences are large, and that optimal allocation is much better; This is also acceptable)

FE6

(c) A preliminary stratified sample was selected from the population with the following results:

| L | T | A |
|---|---|---|
| 10, 1, 17, 9, 20, 14, 6, 22 | 5, 9, 2, 14, 7 | 3, 8, 2, 5 |

Estimate the average # of man-hours lost and place a bound on the error of that estimation.

(d) Estimate the proportion of employees that lost more than one day (8 hours). What is the approximate sample size of proportional allocation required to estimate that proportion with a bound on the error of estimation of 10%?

6 (c)

| | L | T | A |
|---|---|---|---|
| $n_i$ | 8 | 5 | 4 |
| $\bar{y}_i$ | 12.375 | 7.4 | 4.5 |
| $s_i^2$ | 51.696 | 20.3 | 7 |

$$\hat{\mu} = \sum \frac{N_i}{N} \bar{y}_i = \frac{240 \times 12.375 + 130 \times 7.4 + 50 \times 4.5}{420} = 9.898$$

3

$$\hat{Var}(\hat{\mu}) = \sum \left(\frac{N_i}{N}\right)^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i} =$$

$$= \frac{1}{420^2}\left[240^2 \frac{240-8}{240} \frac{51.696}{8} + 130^2 \frac{130-5}{130} \frac{20.3}{5} + 50^2 \frac{50-4}{50} \frac{7}{4}\right] =$$

$$= 2.4365 \qquad B = 2\sqrt{\hat{Var}(\hat{\mu})} = 3.1219$$

3

4/9 FE 7

6 (d) $\hat{p} = \sum \frac{N_i}{N} \hat{p}_i = \frac{1}{420}\left[240 \times \frac{6}{8} + 130 \times \frac{2}{5} + \right.$

$\left. + 50 \times \frac{0}{4}\right] = 0.552 = 55.2\%$ 3

$n = \frac{\sum N_i p_i q_i}{ND + \frac{1}{N}\sum N_i p_i q_i} =$ , use presampling results for $p_i$

$= \frac{240 \times \frac{6}{8} \times \frac{2}{8} + 130 \times \frac{2}{5} \times \frac{3}{5} + 50 \times \frac{0}{4}\ \frac{4}{4}}{420 \times \left(\frac{0.1}{2}\right)^2 + \frac{1}{420}[\ldots \ldots]} = \frac{76.2}{420 \times \left(\frac{0.1}{2}\right)^2 + \frac{76.2}{420}} =$

$= 61.9 = 62$

$\boxed{n = 62}$ 3

FE8

[20] 3) A class list contains 90 students and records the number of summer courses attended by each student. The list is sorted by the number of courses:

| Student | 1 | 2 | ... | 20 | 21 | ... | 56 | 57 | ... | 78 | 79 | ... | 90 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # of courses (y) | 1 | 1 | .... | 1 | 2 | ... | 2 | 3 | ... | 3 | 4 | ... | 4 |

(a) Choose a systematic sample of size n = 15 from the list, assuming random start with student #4. Estimate the average number of summer courses per student. Use some reasonable method to place a bound on the error of that estimation.

(b) Use 4 repeated systematic samples of size 6 to estimate the average number of summer courses per student and place a bound on the error of that estimation. First explain how you would choose this sample, and then use starts with students #2, 6, 8, 13.

6 (a) $N=90$ , $n=15$ , $k=\frac{90}{15}=6$ , sample:

| Stud # | 4 | 10 | 16 | 22 | 28 | 34 | 40 | 46 | 52 | 58 | 64 | 70 | 76 | 82 | 88 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| y | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 |

3 $\bar{y} = \frac{1}{15}(3\times1+6\times2+4\times3+2\times4) = \frac{35}{15} = 2.33$ , $\underline{\hat{\mu} = 2.33}$

Difference method can be applied (monotonic population)

$\Sigma d_i^2 = 3$ , $D^2 = \frac{3}{2\times14} = 0.1071$

$\hat{V}ar(\hat{\mu}) = \frac{N-n}{N}\frac{D^2}{n} = \frac{90-15}{90}\frac{0.1071}{15} = \underline{5.952\times10^{-}}$

3 $\underline{B = 0.154}$



FE9

7(a) $n_s = 6$, $k' = \frac{90}{6} = 15$. There are 15 systematic samples of size 6. Use table of random numbers to select 4 different random numbers between 1 and 15 for sample starts, e.g. 2, 6, 8, 13, and then select these 1-in-15 syst-samples.

| sample | | | | | | | | $\bar{y}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | stud # | 2 | 17 | 32 | 47 | 62 | 77 | |
| | $y$ | 1 | 1 | 2 | 2 | 3 | 3 | $2 = \frac{12}{6}$ |
| 2 | stud # | 6 | 21 | 36 | 51 | 66 | 81 | |
| | $y$ | 1 | 2 | 2 | 2 | 3 | 4 | $2.33 = \frac{14}{6}$ |
| 3 | stud # | 8 | 23 | 38 | 53 | 68 | 83 | |
| | $y$ | 1 | 2 | 2 | 2 | 3 | 4 | $2.33 = \frac{14}{6}$ |
| 4 | stud # | 13 | 28 | 43 | 58 | 73 | 88 | |
| | $y$ | 1 | 2 | 2 | 3 | 3 | 4 | $2.5 = \frac{15}{6}$ |

4

$$\hat{\mu} = \bar{y} = \frac{12+14+14+15}{4 \times 6} = 2.29 = \frac{1}{4} \Sigma \bar{y}_i \quad , n = 6 \times 4 = 24$$

$$s_{\bar{y}}^2 = \frac{1}{n_s - 1} \Sigma (\bar{y}_i - \bar{y})^2 = 0.04398 ,$$

$$\hat{V}ar(\bar{y}) = \frac{k' - n_s}{k'} \frac{s_{\bar{y}}^2}{n_s} = \frac{15-4}{15} \frac{0.04398}{4} = \left( \frac{N-n}{N} \frac{s_y^2}{n_s} \right)$$

$$= 8.063 \times 10^{-3} , \quad B = 0.1796$$

3

FE 10

(c) Do you expect that an SRS of size n = 15 would give better estimate of the population mean than the systematic sample in (a). Justify by some reasonable assumption about intracluster correlation coefficient ρ.

(d) Estimate ρ using some result from (a) and known value of the population variance (calculate the variance from the class list).

3 (c) ordered (increasing) population implies negative $\rho$, and then better estimation from systematic sampling than from SRS.

4 (d) From (a) $\hat{V}ar(\hat{\mu}) = 5.952 \times 10^{-3}$, s.s.n = from the list $\sigma^2 = 0.9165$ (distr. of marks $\frac{1}{20} \Big| \frac{2}{36} \Big| \frac{3}{22} \Big| \frac{4}{12}$)

$$\hat{\rho} = \frac{1}{n-1}\left(\frac{n\hat{V}}{\sigma^2} - 1\right) = \frac{1}{15-1}\left(\frac{15 \times 5.952 \times 10^{-3}}{0.9165} - 1\right) = -0.0645$$

FE 11

[20] 4) A consignment contains 250 boxes of cookies, each containing 20 cookies. Ten boxes were selected at random, all cookies in boxes were checked and weighted, and the following results were obtained:

| Box | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Total weight $y_i$ (gr) | 210 | 195 | 198 | 202 | 205 | 196 | 199 | 200 | 206 | 197 |
| $S_i^2$ | 5 | 3 | 2 | 3 | 4 | 2 | 5 | 8 | 6 | 9 |
| # of cracked cookies | 2 | 1 | 1 | 0 | 0 | 1 | 2 | 3 | 0 | 0 |

(a) Estimate the average weight per cookie in the consignment and place a bound on the error of that estimator.

(b) Estimate the percentage of cracked cookies in the consignment, and place a bound on the error of that estimator.

6 (a) It is one stage cluster sampling

$n=10$, $m=20$, $\bar{\bar{y}}_c = \frac{\Sigma y_i}{n \times m} = \frac{2008}{10 \times 20} = 10.04$

3

$MSB = \frac{m}{n-1} \Sigma(\bar{y}_i - \bar{\bar{y}}_c)^2 = \frac{1}{m} S_y^2 = \frac{23.73}{20} = 1.186$

$\hat{Var}(\bar{\bar{y}}_c) = \frac{N-n}{N} \frac{MSB}{nm} = \frac{250-10}{250} \frac{1.186}{10 \times 20} = 5.696 \times 10^{-3}$

$\hat{SD}(\bar{\bar{y}}_c) = 0.0755$, $B = 0.151$

3

FE 12

6(a) $\hat{p} = \frac{\sum a_i}{n \times m} = \frac{10}{10 \times 20} = 0.05$ 3

$MSB = \frac{1}{m} s_a^2 \quad \left( = \frac{m}{n-1} \sum (\hat{p}_i - \hat{p})^2 \right)$

$= \frac{1.111}{20}$

$\hat{Var}(\hat{p}) = \frac{N-n}{N} \frac{MSB}{n \times m} = \frac{240}{250} \frac{1.111}{10 \times 20^2} = 2.667 \times 10^{-4}$

$\hat{SD}(\hat{p}) = 0.0163, \quad B = 0.0327$ 3

FE 13

(c) Compare the precision of cluster sampling with simple random sampling for the estimation in (a), and decide which design is more efficient.

(d) Estimate the intracluster correlation coefficient ρ from the sample (for weight). Is this in accordance with your result in (~~d~~)? c

5 (c) $MSW = \frac{1}{n} \sum S_i^2 = \frac{1}{10}(5+3+2+\cdots+9) = 4.7$

$\hat{S}^2 = \frac{(M-1)MSW + MSB}{M} = \frac{19 \times 4.7 + 1.186}{20} = 4.52$

$\hat{RE}\left(\frac{\bar{\bar{y}}_c}{\bar{y}_{SRS}}\right) = \frac{\hat{S}^2}{MSB} = \frac{4.52}{1.186} = 3.81$.

Cluster sample is much more efficient.

3 (d) $\hat{\rho} = \frac{MSB - MSW}{(M-1)MSW + MSB} = \frac{1.186 - 4.7}{19 \times 4.7 + 1.186} = -0.039$

$\rho < 0$ implies that cluster sampling is better than SRS, as in (c).

FE/Y

Question 5) ~~##~~ Two stage cluster sample

4 (a) $\hat{\bar{M}} = \frac{1}{n}\sum M_i = \frac{1}{3}(12+16+14) = 14$

$\hat{M} = N\hat{\bar{M}} = 10 \times 14 = \underline{\underline{140}}$ – it is an unbiased estimator, because the sample of plots is an SRS.

(b) use ratio estimator $\hat{\mu}_r = \frac{\sum M_i \bar{y}_i}{\sum M_i}$.

5

| Area | $M_i$ | $m_i$ | $\bar{y}_i$ | $s_i^2$ |
|---|---|---|---|---|
| 1 | 12 | 3 | 3.67 | 2.33 |
| 2 | 16 | 5 | 2.4 | 2.8 |
| 3 | 14 | 4 | 2.5 | 1.67 |

$$\hat{\mu}_r = \frac{12\times 3.67 + 16\times 2.4 + 14\times 2.5}{12+16+14} = \underline{\underline{2.796 = 2.80}} \quad 3$$

$$s_r^2 = \frac{1}{n-1}\sum M_i^2(\bar{y}_i - \hat{\mu}_r)^2 = \frac{1}{3-1}\left[12^2(3.67-2.80)^2 + \cdots\right] =$$

$$= 83.80$$

$$\hat{Var}(\hat{\mu}_r) = \frac{N-n}{N}\frac{1}{n\hat{\bar{M}}^2}s_r^2 + \frac{1}{nN\hat{\bar{M}}^2}\sum M_i^2\frac{M_i - m_i}{M_i}\frac{s_i^2}{m_i} =$$

$$= \frac{10-3}{10}\times\frac{1}{3\times 14^2}\times 83.80 + \frac{1}{3\times 10\times 14^2}\left[12^2\frac{12-3}{12}\frac{2.33}{3} + \right.$$

$$\left. + 16^2\frac{16-5}{16}\frac{2.8}{5} + 14^2\frac{14-4}{14}\frac{1.67}{4}\right] = 0.14073 \quad 3$$

$$B = 2\sqrt{\hat{V}ar} = \underline{\underline{0.750}}$$

FE 15

[20] 5) There are 10 forest areas in the county and they are subdivided into plots. Three areas were selected at random and from each area few plots were selected at random also. The following results are obtained:

| Area | # of plots | # of plots sampled | # of trees in the sample | # of infected trees in the sample |
|---|---|---|---|---|
| 1 | 12 | 3 | 15, 22, 16 | 2, 5, 4 |
| 2 | 16 | 5 | 12, 21, 15, 30, 16 | 1, 3, 1, 5, 2 |
| 3 | 14 | 4 | 18, 10, 10, 16 | 3, 2, 1, 4 |

(a) Estimate the total number of plots in the county. Is this estimator unbiased?

(b) Estimate the average number of infected trees per plot and place a bound on the error of estimation. Is this estimator unbiased?

(c) Estimate the average percentage of infected trees per plot.

(d) Estimate the percentage of infected trees in the county.

---

6 (c) proportions of infected trees per plot in the sample are

| Area | $M_i$ | $m_i$ | $p_i$ | $\bar{p}_i$ |
|---|---|---|---|---|
| 1 | 12 | 3 | $\frac{2}{15}, \frac{5}{22}, \frac{4}{16}$ | 0.204 |
| 2 | 16 | 5 | $\frac{1}{12}, \frac{3}{21}, \frac{1}{15}, \frac{5}{30}, \frac{2}{16}$ | 0.117 |
| 3 | 14 | 4 | $\frac{3}{18}, \frac{2}{10}, \frac{1}{10}, \frac{4}{16}$ | 0.179 |

$$\bar{p} = \frac{\sum M_i \bar{p}_i}{\sum M_i} = \frac{12 \times 0.204 + 16 \times 0.117 + 14 \times 0.179}{12+16+14} =$$

$= 0.163 = 16.3\%$

FE 16

(d) $\hat{p} = \dfrac{\text{\# of infected trees}}{\text{\# of trees}}$

4

$$\hat{p} = \frac{\sum M_i \bar{y}_i}{\sum M_i \bar{x}_i} = \frac{12 \times 3.67 + 16 \times 2.4 + 14 \times 2.5}{12 \times \frac{53}{3} + 16 \times \frac{147}{5} + 14 \times \frac{54}{4}} =$$

$= 0.135 = 13.5\%$

FE17