Last name: First name: Student #:

UNIVERSITY OF TORONTO
Faculty of Arts and Science

DECEMBER 2010 EXAMINATIONS

STA 304H1 F/1003H F

Duration - 3 hours

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

[18] 1) EI Star is a manufacturer of television and other electronic equipment. On the basis of warranty registration cards (filled at purchase), it maintains a file of 21,528 customers who purchased a TV set. From this file, a random sample of 500 customers was selected (without replacement) for the purpose of estimating the proportion of all EI Star customers owning a DVD (regardless of the manufacturer) and the average age of these DVDs. Of the 500 sampled customers, 361 stated they owned a DVD (252 stated having a DVD produced by EI Star). The average age of these 361 DVDs was 14.3 months.
(a) What is the target population? What is the sampling population? What is the frame?
(b) Estimate the number of customers owning a DVD. Is it an unbiased estimator? Explain.
(c) Estimate the average age of all DVDs owned by customers. Is it an unbiased estimator? Explain. **(continued)**

(d) Can you place a bound on the error of estimation in (b), and in (c), using sample data only? Why, or why not? What if it is known that all DVDs are not older than 2 years? Explain.
(e) You want to estimate the percentage of general population of customers owning a DVD. What is the target population? Estimate the percentage from available data from the survey by EI Star. What is the sampling population? Is your estimator biased, considering the purpose of estimation? Discuss it.
(f) Estimate the percentage of general population owning a DVD that owns one produced by EI Star. Is this estimator biased? Discuss at least two sources of possible bias.

[18] 2) For a purpose of planning power production Ontario Hydro selected an SRS of 500 residencies from the population of 2.8 million electricity-using residences in Ontario. Among several other characteristics, the following responses were obtained:

Own a PC: 400; do not own: 100. Total number of PCs in use: 600.
Own vacuum cleaner: 450; do not own: 50. High capacity: 200; regular: 250.

(a) Estimate the average number of PCs in use per residence owning a PC.
(b) Estimate the total number of PCs in use.
(c) Estimate the proportion and number of residences having a high capacity vacuum cleaner among residences having a vacuum cleaner. **(continued)**

(d) Are these estimators in (a), (b), and (c) unbiased? Explain why or why not.
(e) Calculate a bound on the error of estimation of the number of residences having a high capacity vacuum cleaner.
(f) It is known from the sample that 250 residencies own one PC, 100 own two PCs, and 50 own three PCs. Of what particular use this information can be in your study (in what parts of this question)? Explain.

[20] 3) In order to estimate the total inventory of its products being held, a tire company conducts a proportional stratified sample of its dealers, with these dealers being stratified according to their inventory held in previous year. For a total sample size of n = 100, the following data were obtained

Stratum (last year inventory)	Total number of dealers	Sample (current inventory) $\bar{y}_i$	s_i^2
I (0-99)	400	105	400
II (100-199)	1000	180	900
III (200-)	600	282	2500

(a) Find the allocation of the sample used.

(b) From this stratified sample, estimate the mean current inventory μ and the total current inventory.

(c) Estimate the variance of the estimator $\hat{\mu}$ used in (b), and find the 95% CI for μ.

(continued)

(d) What would be the optimal allocation of the sample size n = 100 if the cost of sampling one dealer from stratum III is four times the cost of sampling from stratum I, and the cost of sampling one dealer from stratum II is twice the cost of sampling from stratum I? Use the above given sample as a preliminary survey.

(e) Ignoring the sampling costs, do you expect that the stratified sample with the proportional allocation will be significantly better than an SRS? What if you compare the optimal allocation and the proportional allocation? Explain, but don't calculate the variances.

[12] 4) Assume that 20 households listed below make a village, presented in their order along the main road.

household	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Males (y)	3	2	3	2	2	2	1	5	3	2	2	2	5	4	2	2	4	3	1	2
Fem. (x)	3	2	2	2	2	2	2	4	2	3	3	2	3	3	1	2	3	5	0	4

(a) Calculate **the theoretical** standard deviations of (i) the sample mean of the household sizes from an SRS of 5 households, and of (ii) the sample mean of the household sizes from a systematic sample of 5 households from the list. (iii) Which sample, SRS or systematic is expected to give a more efficient result?

(b) If the road were much longer and with more households would you expect similar results from an SRS, as from a systematic sample of the same size? Explain.

(c) If you select an SRS of households from the road and record the household size and the house age (in years) for each household, would using a ratio (or regression) estimator be better than just using an SRS estimator? Explain. Assume that the average age of houses is known.

[15] 5) A city transportation system includes 30 subway stations, each containing 6 escalators operating, and is interested in the number of days that the escalators were down for repair in the past year. Assume that 3 subway stations were selected at random into the sample, and the maintenance records for all 6 escalators were examined. The following results were obtained.

station	days escalators down	average	variance
1	4 3 7 2 11 0	4.5	15.5
2	11 4 3 1 0 2	3.5	15.5
3	0 3 6 4 3 2	3	4

(a) (i) Explain what kind of sampling design is used here. (ii) Estimate the average downtime per escalator and place a bound on the error of that estimation.
(b) Estimate the total numbers of days down of all escalators in the subway and place a bound on the error of that estimation. **(continued)**

(c) Compare the efficiency of cluster sampling with that of SRS in this case (use sampling results only) and decide which sample design is more efficient.
(d) Estimate the intracluster correlation coefficient from the sample. Is this estimate in accordance with your result in (c)? Explain.

[17] 6) A bank has 10 branches around the city. The 1995 profits (x), 1999 profits (y) (in units of $1,000,000) and the number of employees in 1999 (m) were as follows:

branch	1	2	3	4	5	6	7	8	9	10
m	4	2	3	2	6	2	2	5	7	4
x	6	4	6	2	14	5	4	11	15	8
y	8	5	5	3	17	4	5	13	14	10

(a) Calculate the theoretical variance of the sample mean for the 1999 profit from an SRS of size 4 from these 10 branches.

(b) Explain in detail how you would select a PPS sample of size n with probability proportional to size of employees. Select a sample of size 4 using the following portion of a table of random numbers.

01224 76384 97403 53363 44163 64486 64758 75366 76554 31601 12614 33072
19474 23632 27889 47914 02584 37680 20801 72152 39339 34086 43218 15263.

(continued)

(c) Assume the branches # 2, 4, 5, 9 were selected into sample in (b). Estimate the average 1999 profit and the variance of the estimator. Compare this result with one in (a), and give a comment on it.

(d) A PPS sample with probabilities proportional to size of 1995 profit could be also used. Discuss in short which sampling design, one from (a), from (b), or this one from (d) would be the best one (in principle) for estimation of the average 1999 profit.

Total marks = 100