Last name: First name: Student #:

UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL/MAY 2013 EXAMINATIONS

STA 304H1 S/1003H S

Duration - 3 hours

Examination Aids: Non-Programmable Calculator; aid-sheet, one sided, or two sided, with theoretical formulas only, as posted on the web-site. You may use any back side, with clear indication of Question part.

Question	1	2	3	4	5	6	Total
Max mark	21	18	21	12	18	10	100
Mark							

Q. 1. [21]
A tourist company organizes various types of sight tours, offering some discount rates for certain types of customers, among them children (under 12) and seniors (over 65 years of age). The company wants to estimate the total number of seniors on its tours on the basis of a random sample of site tours recorded in their books. Each tour record shows the number of seniors, the number of children, and some other data. There were 4,000 tours organized during the last year, listed as they appeared in calendar time. The records show seasonal variations in numbers and types of people on the tours. Most of the tours were organized in summer and least in winter.
(a) Describe briefly how you would select an SRS of 100 tours from the last year.
(b) The company's manager prefers a simpler method, using one-in-40 systematic sampling of tours. Describe briefly how you would select this sample.
(c) Can the sample obtained in (b) be treated as an SRS for the purpose of estimation? Explain why, or why not. **(continued)**

(d) The sample in (a) was selected and the average number of seniors per tour was 20 seniors, with the sampling standard deviation of 5 seniors. Estimate the total number of seniors taking tours last year, and place a bound on the error of estimation.

(e) The company also wants to estimate the percentage of seniors out of all tourists on the tours last year. (i) What information should be included in the sample to be able to estimate this percentage, and to place a bound on the error of estimation? (ii) Exactly what values should be calculated from the sample to be able to complete the required tasks?

(f) It is estimated that the total number of tours next year will be 10% higher than this year. How large should next year's SRS sample be so that the estimated total number of seniors will be within $\pm 2{,}000$ of the true total with probability 95%?

(g) What sampling design would you suggest that you think would produce better results than one in (a) or (b)? Explain your choice in some details.

Q. 2. [18]

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. (some residencies may own several PCs)
>
> Own electric stove: 450; do not own electric stove: 50 (likely, they use gas). Glass-ceramic cooktop: 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 glass-ceramic stove among residences having an electric stove. **(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 Glass-ceramic cooktop stove.

(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 Question 2)? Explain and use it only in one part, if it can be used in more than one.

Q. 3. [21]

A market research firm conducted a survey in 2000 in a city for the purpose of estimating the total monthly household expenditures on compact discs (CDs) and the total number of households owning a compact disc player (CDP). The city was divided into four geographical areas and a random sample of households was selected from each area. The results of the survey are as follows:

Area	Number of Households	Number Sampled	Sample Average Monthly Expenditure ($)	Sample Proportion Owning a CDP
A1	20,000	100	20.80	30%
A2	10,000	100	12.20	16%
B1	35,000	100	8.10	8%
B2	15,000	100	16.48	14%
Total	80,000	400		

(a) (i) Estimate the average monthly household expenditure on CDs in the city, and (ii) the proportion of households in the city owning a CDP.

(b) (i) Estimate the total monthly household expenditure on CDs in the city, and (ii) the total number of households owning a CDP in the city.

(c) (i) How many households would be sampled from each area if the sample of 400 were with proportional allocation? (ii) Considering already obtained sample, do you think that the stratified sample with proportional allocation would produce better results than an SRS of the same size, for the both parameters in (a), for only one of them, or none of them?

(continued)

(d) Explain on which parameter estimated in (a) you can place a bound on the error of estimation, and then calculate it.
(e) (i) Can you find the optimal allocation for estimation of the proportion of households in the city that own a CDP? Explain, but don't calculate anything. (ii) Would this allocation be likely near to optimal for estimation of the average monthly household expenditure on CDs? Explain.

Q. 4. [12]

From a directory of 16 households on a street, the actual numbers of people living in the households (household size) are as follows:

Household															
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
5	5	4	2	5	4	5	4	3	4	4	3	2	6	1	3

(a) Calculate **the theoretical** standard deviations (i) of a systematic sample of one in four households from the directory and (ii) of an SRS of the same size for estimating the average household size. Which design is expected to give better result?

(b) If the street were much longer would you expect same results from an SRS, as from a systematic sample of the same size? Explain.

(c) If you select an SRS of households from the street 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 all houses is known.

Q. 5. [18]

A psychologist wants to determine how many hours per week 8-10 year old boys in Toronto spend playing video and computer games. To investigate this she randomly selects three Toronto primary schools from the population of 230 primary schools in Toronto, and then randomly selects 20 (twenty) 8-10 year old boys from each school. With the parents help, she has each of the sampled boys record their video and computer game times for a calendar week, in hours. The results are presented in the following table:

School	Number of 8-10 year old boys	Sample size	Sample Mean	Sample St. Dev.
A	350	20	15	5
B	200	20	12	4
C	150	20	11	3

(a) Explain what kind of design is used here. Do you think this design is appropriate? Explain (ignore the small sample size problem).

(b) Comment on the condition of the design "... *for a calendar week* ...". Would the selection of a "calendar week" affect results of the study?

(c) (i) Estimate the total number of hours spent by 8-10 year old Toronto boys during a week on video and computer games, and (ii) the variance of the estimator (you do not need to complete the calculation of the variance, but you have to calculate all entries). (iii) Is the estimator in (i) unbiased? Explain. **(continued)**

(d) Do you expect that your standard error in (c) would be larger, smaller or about the same as the standard error of an estimator based on SRS of the same size (60)? Explain.
(e) (i) Estimate the percentage of the free time of 8-10 year old boys spent during a week on playing the games. Assume that 8h a day is spent on sleep (7 days a week), and 7h in school and transportation (5 days a week). (ii) What kind of estimator are you using in (i)? Can you place a bound on the error of this estimator? Explain, but don't calculate, if you can.

Q. 6. [10] ("theoretical" question)
In Question 2, an SRS of 500 residencies from the population of 2.8 million electricity-using residences in Ontario was selected. Among several other characteristics, the following responses were obtained:

Own electric stove: 450; do not own electric stove: 50. Glass-ceramic cooktop: 200; regular: 250.

You estimated in (c) the proportion of residences having a glass-ceramic stove among residences having an electric stove. This proportion applies, obviously, only to residencies with an electric stove. Consider now the general case:

Given: N – population size, n –sample size of an SRS of residencies,
n_1 – the number of residencies in the sample having an electric stove,
n_2 – the number of residencies in the sample having a glass-ceramic electric stove ($n_2 \leq n_1$), $\hat{p}_1 = \frac{n_1}{n}$, $\hat{p}_2 = \frac{n_2}{n_1}$.

(a) $\hat{p}_2$ is our estimator for the proportion of residences having a glass-ceramic stove among residences having an electric stove. Derive a convenient general formula for $\hat{V}ar(\hat{p}_2)$ (estimated $Var(\hat{p}_2)$) using the above notation, where needed. (hint: what kind of estimator are you using?)

(b) Calculate $\hat{V}ar(\hat{p}_2)$ from the actual sample data using your formula.

///an extra page for work