Last name: First name: Student #:

UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL/MAY 2007 EXAMINATIONS

STA 322H1 S

Duration - 3 hours

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

[14] 1) A marketing research firm (MRF) estimates the proportion of customers preferring a certain brand of lipstick by "randomly" selecting 100 women who came by their booth in a shopping mall. Of the 100 sampled, 65 women stated a preference for brand A.
(a) Assuming all is correct with this sampling method, estimate the true proportion of women preferring brand A, and place a bound on the error of estimation.
(b) If a more accurate estimate has to be found, propose the minimal sample size required to estimate the true proportion with the bound of 3%. A reasonable guess is that the true proportion is within ±10% of the value obtained in (a).
(c) What is the target population in this study? What is the sampling population in this study? **(continued)**

(d) Did MRF select a simple random sample from the target population? Explain. Even if the sample may not be an SRS, could it still be reasonably representative of the target population? (consider there is a number of shopping malls in every city)
(e) How would you help MRF to improve their sampling strategy? (consider discussion in (c) and (d)).
(f) Does sampling population consisting of women visiting shopping malls invokes a significant bias in the study? Give arguments for , and against. If MRF wants to extend its sampling population, what method of data collection can it use? Would it significantly increase the cost of the study, compared to the method already used?

[16] 2) From a small suburban community of 24 households, 5 were selected at random, and the number of adults, number or children, and total daily income were recorded for each.

Household	1	2	3	4	5
Number of adults, y	3	4	3	5	6
Number of children, z	2	4	1	4	4
Total daily income, x (in $10)	33	40	34	68	61

($\sum y=21$, $\sum y^2=95$, $\sum x=236$, $\sum x^2=12190$, $\sum xy=1067$, $\sum z=15$, $\sum z^2=53$, $\sum zy=69$)

(a) Estimate (i) the total number of residents in the community, (ii) total number of children, (iii) average household size, and (iv) total daily income in the community.

(b) Estimate (i) the average monthly income per household and per adult (one month = 30 days), (ii) average number of children per household, (iii) percentage of children in the community population, and (iv) average percentage of children per household.**(continued)**

(a) What kinds of estimators are used in (b) (answer for each case)?

(b) (i) Place an error bound on the estimator of the average monthly income per adult in (b)(i). (ii) Place an error bound on the estimator in (b)(iv).

[20] 3) The households in the community from Q. 2 (24 in total) are divided into two strata, by household size, stratum I with households of size 1-5 persons (17 households), and stratum II with households of size 6 and more (7 households). A stratified random sample of size 5 was selected from the community, and the data on daily income and daily food cost were recorded, as follows:

Household	Stratum	Size	Daily income, y ($10)	Daily food cost, x ($10)
1	1	3	33	21
2	1	4	39	24
3	1	3	35	22
4	2	6	62	31
5	2	6	61	29

(a) Estimate the total size of the community, total daily income in the community, and total daily food cost.
(b) Place a bound on the error of estimation of the total daily food cost.
(c) Estimate the average daily income per person, and percentage of income spent on food in the community. Are these estimators biased, or unbiased? Explain. **(continued)**

(d) Estimate the average percentage of income spent on food per household. Is this estimator biased, or unbiased? Explain.
(e) Would a stratified sample with proportional allocation be better in this survey than an SRS? (ignore that the population is very small) Explain.
(f) (i) Is this type of stratification good, or not good for this survey on expenditures? Explain. (ii) Would an optimal allocation of the sample produce better results than a proportional in this survey? (ignore that the population is very small) Explain.

[14] 4) In a recent review of the largest Canadian Tech companies (The Globe and Mail, April 2007), 250 companies are listed in descending order of their 2006 revenues (in $1000). The first 10 companies have revenues significantly larger than the rest, totaling 40,126,556. The other 240 companies have revenue between 1,000, 000 and 0 (the smallest revenue is 3,050). To estimate the total revenue of all tech companies, I selected 5 different systematic samples of size 10, with a random start, from the remaining 240 companies. The following results were obtained:

Sample and start	1: 13	2: 20	3: 28	4: 30	5: 16
Total revenue in the sample	847,991	558,491	426,927	207,080	477,650

(a) (i) Explain why using a systematic sampling may be a good idea in this estimation. (ii) What is the selection step? (iii) Is it by chance that the total revenue in sample 1 is almost twice larger that in sample 3? Explain.

(b) (i) Estimate the total revenue for all 250 companies. (ii) Estimate the average revenue for first 10 companies, and (iii) for remaining 240 companies. **(continued)**

(c) Estimate the standard deviations of the estimators used in (b)(ii) and (iii) (be careful here).
(d) Would it be better to use an SRS of size 50 from 240 lower companies, to estimate their average revenue, than to repeat 5 times systematic samples of size 10, as it was done above? (hint: try to estimate the theoretical variance of the sample mean of an SRS of size 50 from the population)
(e) (bonus) Assuming that only one systematic sample of size 50 was selected from the population of 240 lower companies, how would you estimate the variance of the sample mean? Try to use some reasonable guess and do the estimation.

[20] 5) A convenience store company conducts an investigation on its franchises in the 500 places in the country. Eight places were selected at random, and all franchises from each place were examined. The following results were obtained:

Place	1	2	3	4	5	6	7	8
Number of franchises (x_i)	8	12	4	5	6	6	7	5
Number of franchises in shopping malls (y_i)	4	7	1	3	3	4	4	2
Total daily revenue in selected franchises ($1000)	8	10	5	4.5	7	4.5	8	4

$\Sigma x_i = 53$, $\Sigma x_i^2 = 395$, $\Sigma y_i = 28$, $\Sigma y_i^2 = 120$.

(a) Explain what kind of sampling design is used here. Estimate the total number of these franchises in the country and place a bound on the error of estimation.
(b) Estimate the total number of franchises in malls in the country.
(c) Place a bound on the error of estimation on the proportion of franchises in malls in the country. **(continued)**

(d) Estimate the total daily revenue of these franchises in the country? Can you place a bound on the error of estimation? Just explain, don't calculate.
(e) What other parameters of interest besides already mentioned in (a)-(d) is possible to estimate from this data? Name at least two (you don't need to estimate them).
(f) It is known that the total number of these franchises in the country is 3300. Using that information, estimate the total number of franchises in malls in the country. Which method of estimation would you prefer, one from (b) or this one from (f)? Justify, without calculating variances.

[16] 6) In a study of a particular forest disease, an affected county was selected for a more detailed investigation. The forest in the county is divided into 10 forest areas, and further each area is divided into roughly equal plots. Three areas were selected at random, and then a few plots from each of selected plots. For each plot, the number of trees, and the number of infected trees were counted. The following results were obtained:

Area	Number of plots	Number of plots sampled	Number of trees (y) and number of infected trees (x) in the sampled plots						Mean	Variance
1	12	3	x	2	5	4			3.67	2.33
			y	15	22	16				
2	16	5	x	1	3	1	5	2	2.40	2.80
			y	12	21	15	30	16		
3	14	4	x	3	2	1	4		2.50	1.67
			y	18	10	10	16			

(a) What kind of sampling design is used here? What is the target population in this survey? What is the actual sampling population? Estimate the total number of plots in the county.

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

(c) Estimate the total number of infected trees per area. Is this estimator unbiased?

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

Total marks = 100