UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2009 EXAMINATIONS
STA 304H1F / STA 1003H F

Duration - 2 hours

Examination Aids: one 2-sided handwritten aid sheet (8 ½ x 11)
and non-programmable calculator.

Answer all questions in the examination book.

1. **(30 marks)** You have been hired to estimate the average weekly entertainment expenses for students at the University of Toronto. You have estimated that you have resources to interview approximately 60 students.

   (a) Describe how you will choose the target population and how you might obtain an appropriate sampling frame.

   (b) Describe, with specific details, how you would obtain a sample of 60 students by simple random sampling, by stratified sampling, and by cluster sampling. What are some advantages and disadvantages of each method?

   (c) Your competition has decided to carry out two stage cluster sampling. There are 100 classes that meet on Monday morning at 10 am, and they will use these classes as their psu's. They have decided to take a 10% sample of students in each psu and interview them about their monthly expenses. Suggest two potential biases that this proposal might have, and explain what might be done to reduce or eliminate these biases.

   (d) Your study, using a two-stage cluster sampling design, has been completed, and the results are shown in Table 1 on the next page. You may assume that potential biases have been effectively eliminated. Provide a 95% confidence interval for the average weekly entertainment expenses among students.

   (e) Summarize the results of your study in a short paragraph. Describe briefly, in non-technical language, the survey, the results, your conclusions, and any limitations.

Table 1: Summary data for Question 1. Six classes were chosen from 100 psu's.

Number of students in each class, M_i	Number of students sampled, m_i	$\bar{y}_i$	s_i^2	$M_i\bar{y}_i$	$\{M_i(\bar{y}_i - \hat{\bar{y}}_r)\}^2$
150	15	25	13	3750	317,678
80	8	16	9	1280	1,041,629
47	5	30	15	1410	3,410
62	6	21	22	1302	231,329
39	4	45	18	1755	401,267
220	22	35	17	7700	1,886,072
598				17197	3,881,386

$$\sum_{i=1}^{6}(1 - \frac{m_i}{M_i})M_i^2\frac{s_i^2}{m_i} = 82,485.2$$

2. (**20 marks**) Investigators selected a simple random sample without replacement (SRS) of 200 high school students in Grade 12, from a population of size 2000, for a survey of TV-viewing habits, with an overall response rate of 75%. By checking school records, they were able to find the grade point average (GPA) for the non-respondents and classify the sample accordingly:

	Sample	Number of	Hours of TV	
GPA	Size	Respondents	$\bar{y}$	s_y
3.00 – 4.00	75	66	32	15
2.00 – 2.99	72	58	41	19
below 2.00	53	26	54	25
Total	200	150		

$$\{(32 \times 66) + (41 \times 58) + (54 \times 26)\}/150 = 39.3$$
$$\{(15^2 \times 65) + (19^2 \times 57) + (25^2 \times 25)\}/149 = 341.1 = (18.5)^2$$

(a) What is the estimate for the average number of hours of TV watched per week if only respondents are analyzed? What is the standard error of the estimate?

(b) A χ^2 test of the hypothesis that the response rates are constant across the three GPA groups gives a p-value of 0.012. What does this say about the type of non-response: is it likely to be MCAR (Missing Completely at Random), MAR (Missing at Random), or Non-ignorable?

(c) You've been told that the goal is to use the GPA classification should be used to adjust the weights of the respondents in the sample. What is this method called?

(d) What other methods might you use to adjust for the non-response?

3. **(15 marks)** Suppose a three-stage cluster sample is taken from a population with N psu's, M_i ssu's in the ith psu, and L_{ij} tsu's (*tertiary sampling units*) in the jth ssu of the ith psu. To draw the sample, n psu's are randomly selected, then m_i ssu's from the selected psu's, and then ℓ_{ij} tsu's from the selected ssu's, using SRS at each stage.

   (a) Show that the sample weights are

$$w_{ijk} = \frac{N}{n}\frac{M_i}{m_i}\frac{L_{ij}}{\ell_{ij}}.$$

   (b) Let

$$\hat{t} = \sum_{i\in S}\sum_{j\in\mathcal{S}_{\rangle}}\sum_{k\in\mathcal{S}_{\rangle|}} w_{ijk}y_{ijk}.$$

   Show that $\hat{t}$ is unbiased for $t = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{L_{ij}} y_{ijk}$.

   (c) Describe, without formulas, how you would assess the sampling variance of $\hat{t}$.

4. **(15 marks)** Define **any four** of the following terms, and illustrate each with an example:
(a) optimal allocation; (b) random digit dialling; (c) systematic sampling;
(e) questionnaire bias; (e) ratio estimation; (f) sampling frame;
(g) survey weights

5. **(20 marks)** A researcher wants to estimate the total number of acres planted in icewine, in wineries in Southern Ontario. Because the number of acres varies considerably with the size of the winery, the 75 wineries in the region are placed in one of four categories, according to size, and 15 individual wineries are selected by SRS within these size groupings. The results are given in Table 2.

   (a) Estimate the total number of acres in icewine in Southern Ontario.

   (b) What sampling method was used for this survey? What drawbacks does it have?

   (c) Suggest an alternative survey design and explain why you think it might be preferred.

Table 2: Results from a sample survey of the 75 wineries in Southern Ontario.

Total acreage of winery	Number of wineries	Number sampled	Number of acres in ice-wine	Sample total	Sample s^2
0 – 20	35	7	11, 14, 7, 13, 10, 11, 11	78	5
20 – 40	20	4	13, 14, 19, 15	61	5.5
40 – 60	10	2	22, 30	52	27
more than 60	5	1	40	40	0