Last name, first name: . Student #:

STA 322H1 F SUMMER 2008, First Test, May 29 (20%)
Duration: 50min. Allowed: nonprogrammable hand-calculator, aid-sheet, one side, with theoretical formulas and definitions only. Test consists of 5 pages.

[40] 1) A large medical professional organization with a membership consisting of doctors, nurses and other medical employees wanted to know how its members felt about the organization. The organization also has regional branches that conduct regional affairs. The organization has its central register, and also the regional registers. Each register includes the name, membership type (doctor, nurse, other), and place of living, among other information. The registers are updated every two years.

(a) What is the target population in this study?

(b) If the organization wanted to select an SRS of its members and send them a questionnaire, what can be used as a frame? Discuss, in short, possible problems with this frame.

(c) A pretest (presample) is conducted in (b) to estimate the proportion of members willing to participate in the survey. What should be the sample size of the SRS so that the above proportion can be estimated with a bound on the error of estimation of 3%? From the past experience it is known that this proportion is not greater than 25%.

(continued)

Solutions:
[5] (a) All current members of the organization – doctors, nurses and others.

[6] (b) The central register can be used as a frame (not regional registers). **[2]**
Problem is that the register is updated every two years, so that the frame does not cover the population properly – new members are missing, and some old are not the members anymore. **[4]**

[8] (c) Using large N, $n = \frac{Npq}{(N-1)D + pq} \approx \frac{pq}{D}$, **[2]**

$D = (B_p / 2)^2 = (0.03/2)^2 = (0.015)^2$, **[2]**

$pq \le 0.25 \times 0.75 = 0.1875$, **[2]**

$n \le \frac{0.1875}{(0.015)^2} = 833.3$. So, n = 834. **[2]**

(d) If the organization wanted to select 2% of its members in (b), what would be the simplest (fastest) way to do it? Could this sample be considered as an SRS? Explain.

(e) Consider the following design: The organization selected 500 members from each of the membership types, and then surveyed those 1500 members. (i) Name the type of the sampling plan used. (ii) Give at least one specific goal of the study when this sampling plan is appropriate. (iii) Is selecting 500 members from each group a smart decision?

(f) Consider the following design: The organization randomly selected 20 places (using SRS) from all places in which its members lived, then surveyed 20% of members in those places. (i) Name the type of the sampling plan used. (ii) Suggest at least one problem with this design, regarding the overall goal of the study and the accuracy.

Solutions:
[7] (d) The simplest way would be to use systematic sampling, by selecting every 50^{th} member from the register using random start. **[4]** This sample can be considered as an SRS if the members are in an order unrelated to the study, e.g. by last names (alphabetic order). **[2]** If the members are ordered, e.g., by membership type, the sample cannot be considered as an SRS. **[1]**

[8] (e) (i) Stratified sampling. **[4]** (ii) This would be appropriate if we want to investigate how members of each type feel about the organization. **[2]** (iii) It might not be a good idea, because these groups differ in size, and this selection may underrepresent, or overrepresent some of them. **[2]** (Selecting the sample sizes from each group may depend on other factors, too.)

[6] (f) (i) Two-stage random sampling. **[4]** (ii) The problem is that all places (small and large) have the same chance of being selected, but the large places (such as Toronto) represent a significant portion of the membership. **[2]**

[45] 2) We are interested in investigating the number and type of the summer courses taken by the students registered in the STA322 summer course. The following diagram is available. In the diagram, a classroom with 90 seats and students from the course, as occurred on one of the lectures, is presented. Every rectangle represents a seat, and the accompanying number is the number of the summer courses that the student seating on it is currently registered for (23 shaded seats are unoccupied).

Front

1	2	3	4	5		6	7	8	9	10
	1	2	2	1	1	3	1	2	1	
	2	1	1	1	2	2	1	1	2	
1	2	3	2	1	3	1	1	2	1	1
2	1	1	1		4	3	1	2	2	
	1	2	1	1	5	1	1	2	3	1
2	1	1	1	2	6		1		1	
	2	3	1		7	2	1	1	1	2
1	1	2	1	2	8	3	2			
	1				9	1				

(a) What is the target population in this study? What is the frame? What are the sampling units? Is there any problem with the frame? Explain.
(b) Using the following portion of the table of random numbers, select an SRS of size 10 from the class. Explain your procedure of using the table, and show the procedure. Present the sample that you have obtained (variable values!).
31624 76384 17403 53363 44167 64486 64758 75366 76554 31601 12614 33072 60332 92325 19474 23632 27889 47914 02584 37680 20801 72152 39339 34806 08930 25570

(continued)

Solutions:
[6] (a) Target population: All the students registered in the STA322 summer course. **[1]**
Frame: Classroom plan. **[1]**
Sampling units: Occupied seats (seats with students), or students present in the classroom. **[2]**
Problems: Some registered students are not present, and some unregistered students may be present. **[2]**

Solutions**:**

(b) The population size is smaller than 100, so two random digits should be used, e.g., by reading consecutive pairs without omitting any, starting from the beginning. Let the first digit represent row, and the second column, where 10^{th} column can be represented by 0, so the used groups are
11, 12, 13,…, 19, 10; 21, 22,…, 29, 20;…; 91, 92,…, 99, 90. Unused groups are 01,…, 09, 00. Unoccupied seats are ignored (any other proper representation of the sampling units is acceptable, if explained). Using digits 31624 76384 17403 53363 44167, the result is

digits	31	62	47	63	84	17	40	35	33	63	44	16
unit	31	62	47	63	84	17	free	35	33	Rep.	44	16
variable	1	1	1	1	1	1		1	3		1	3

Sample obtained: 1, 1, 1, 1, 1, 1, 1, 3, 1, 3.

(c) Assume the following sample was obtained: 2, 1, 3, 2, 1, 1, 2, 1 (ignore your sample). Estimate the average number of summer courses taken by the students, and calculate a bound on the error of estimation.
(d) Using the sample from (c), estimate the proportion of students taking exactly two courses, and find the confidence interval on the proportion.
(e) Assuming that there are 75 students actually registered in the course, estimate the total number of summer courses taken by the class. Estimate the variance of this estimator (be careful here).
(f) **(bonus)** Assuming there were no curious students present in the class (ones not registered for the course), are our estimates of the average and total for the class (target population) unbiased or slightly biased? Explain. If you think they are biased, can you argue that they will underestimate, or overestimate the true values? Explain.

Solutions:
[10] (c) $\hat{\mu} = \bar{y} = \frac{1}{8}\Sigma y_i = \frac{1}{8}(2+1+\cdots+1) = 1.625$, **[5]** $S^2 = \frac{1}{7}\Sigma(y_i - \bar{y})^2 = 0.55357$,
$\hat{Var}(\hat{\mu}) = \frac{67-8}{67}\frac{S^2}{8} = 0.08190$, $B_\mu = 2\times\sqrt{\hat{Var}(\hat{\mu})} = 0.572$. **[5]**

[10] (d) $\hat{p} = \frac{3}{8} = 0.375$, **[4]** $\hat{Var}(\hat{p}) = \frac{67-8}{67}\frac{\hat{p}\hat{q}}{8-1} = 0.029484$, $B_p = 2\times\sqrt{\hat{Var}(\hat{p})} = 0.3434$,
$CI_p = \hat{p} \pm B_p = [0.375-0.343, 0.375+0.343] = [0.032, 0.718]$**. [6]**

[9] (e) $\hat{\tau} = N_{reg.class} \times \bar{y} = 75\times 1.625 = 121.875$, **[5]**
$\hat{Var}(\hat{\tau}) = \hat{Var}(N_{reg.class}\hat{\mu}) = N^2_{reg.class}\hat{Var}(\hat{\mu}) = 75^2 \times 0.08190 = 460.69$. **[4]**

[5] (d) (bonus) We may argue that our estimators for the mean and total are slightly biased, and give underestimation because it may be expected that the student with more than one registered course are more likely to be absent from the class (more tired/ more busy) than the students with one registered course (only this one). Somebody may argue that the students with more than one registered course are more ambitious, and then more likely to be present, which I doubt, but this can be accepted as an answer, also. In that case the estimators would give slight overestimation. (Due to small number of students with 3 courses, a small bias may be expected.) **[5] (only a good answer is accepted, either 0 or 5 points)**

[15] 3) Assume that the frame in the previous problem represents the population properly.

(a) Find the population distribution of the number of registered courses per student.

(b) Calculate the population mean and the variance from the distribution.

(c) What are the errors of estimation in questions 2) (c) and 2) (d)?

Solutions:

[5] (a) Distribution of y = number of courses taken by a student (from the class)

y_i	1	2	3	total
n_i	39	22	6	67

[6] (b) $\mu_y = \frac{1}{67}(1\times 39 + 2\times 22 + 3\times 6) = \frac{101}{67} = 1.507$, **[3]**

$\sigma^2 = \frac{1}{67}(1^2\times 39 + 2^2\times 22 + 3^2\times 6) - \left(\frac{101}{67}\right)^2 = 0.429$.**[3]**

[4] (c) In 2) (c) $|\hat{\mu} - \mu| = |\frac{13}{8} - 1.507| = 0.118$, **[2]**
and in 2) (d) $|\hat{p} - p| = |\frac{3}{8} - \frac{22}{67}| = 0.0466$. **[2]**