Last name, first name: . Student #:

STA 304H1 F, FALL 2010, First Test, October 22 (20%)
Duration: 1h. Allowed: nonprogammable hand-calculator, aid sheet, one side, with theoretical formulas and definitions only.

[34] 1) A survey is conducted in a cottage country place in Muskoka region with a goal to investigate the age structure, occupation, and some other characteristics of the residents. The place has 1000 households. Explain in more detail some of the basic concepts in this survey:

(a) [3] Population of interest (target population).
(b) [3] Population parameters of interest (name a few).
(c) [3] Sample design (one you think might be appropriate).
(d) [3] Sampling units and frame.
(e) [5] Variables of interest.
(f) Method of data collection (list the most common methods of data collection and propose one you think might be appropriate here) [5]
(g) [5] Would time of data collection (during the year) matter in this survey? Explain why or why not?
(h) [7] The total number of residents in the place should be estimated with an error bound of 100 residents. Give an estimate of the required sample size, assuming there will be no nonresponse (the survey will collect data on the number of residents in each household in the sample).

Solution:
(a) All residents in the place. Some people might not live permanently in the place (use as a cottage only), so the population of interest might be only the people who live permanently in the place (both answers are acceptable). **[3]**
(b) Proportions of age groups, proportion of children, seniors etc, average age in various groups, such as males, females, permanent residents, etc.**[3]**
(c) Two state cluster sample would be simple to obtain, i.e. first an SRS of households, then SRS of residents from each household, or just one-stage cluster sample of households, and then all residents in the household. If a list of all residents is available, an SRS of residents may be used. For the goal of the survey, the cluster sampling might be more appropriate and simpler to select.**[3]**
(d) For cluster sampling: Sampling units - households, frame - list of households. For SRS: Sampling units - residents, frame - list of residents (if available). **[5]**
(e) For sampling of households: family size, age of family members, number of children, their gender (sex), ...
For sampling of persons: age, sex, family status (single, have family), ... **[5]**
(f) Direct observations, personal interviews, telephone interviews, mailed questionnaires. **[3]** In this study: personal interviews, i.e. visits of trained interviewers, or by sending self-administered questionnaires. **[2]**
(g) Summer time would be better, because the nonpermanent residents would be more likely present, so the possible nonresponse problem (residents are not in the household) would be reduced. The worst time would be winter, regarding the nonrespose problem. If it were a ski resort, a winter time might be better. **[5]**

(h) When households are sampled, $n = \frac{Ns^2}{(N-1)D+s^2}$, where N = 1000, $D = \left(\frac{B}{2N}\right)^2 = \left(\frac{100}{2000}\right)^2 = \frac{1}{400}$ **[2];** assuming that the number of residents in a household is between 1 and 5, using range, $s = \frac{R}{4} = \frac{5-1}{4} = 1$. Then $n = \frac{1000\times1}{999/400+1} = 286$. **[5]** (any similar range for the number of residents is acceptable)

[46] 2) In a study of people residing in a large city, a simple random sample of 1203 households was selected, and the men over 30 were counted. The sample contained a total of 1073 men over 30 years old, distributed as follows:

# of men over 30	0	1	2	3	4
# of households	232	876	88	7	0

(a) [6] Do the men over 30 in the sample comprise a simple random sample of all such men in the city? Why or why not?
(b) [10] Make a 95% confidence statement on the proportion of all households in the city containing men over 30 years of age.
(c) [10] Estimate the average number of men per household in the city, and place a bound on the error of estimation.
(d) [8] Assuming there are 60,000 households in the city, estimate the total number of men over 30 in the city.
(e) [12] Assuming same as in (d) and also that there are 75,000 men in the city, estimate the proportion of men over 30 of all men, and place a bound on the error of the estimation. (be careful here, there is a catch in the last part; if you don't see it, better go to the next question)

Solutions:
(a) No, the sample consists of random groups of men (households) rather than random individual men, i.e. each man is not equally likely of being selected. **[6]**

(b) $\hat{p} = (876 + 88 + 77 + 0)/1203 = 971/1203 = 0.807$. **[5]**
Assuming (N - n)/N ≈ 1, $B_p = 2\sqrt{\hat{p}\hat{q}/(n-1)} = 2\sqrt{0.807\times0.193/1202} = 0.023$,
$I_p = [0.807 - 0.023, 0.807 + 0.023] = [0.784, 0.830]$. **[5]**

(c) $\hat{\mu} = \bar{y} = \frac{1}{1203}(0\times232 + 1\times876 + 2\times88 + 3\times7) = \frac{1073}{1203} = 0.8919$, **[5]**
$S^2 = (232\times(0-0.892)^2 + 876\times(1-0.892)^2 + \cdots)/(1203-1) = 0.278$ **, [2]**
$\hat{Var}(\hat{\mu}) = \frac{N-n}{N}\frac{S^2}{n} \approx \frac{S^2}{n} = \frac{0.278}{1203} = 2.31\times10^{-4}$ **,**
$B_\mu = 2\times\sqrt{\hat{Var}(\hat{\mu})} = 0.03$**. [3]**

(d) $\hat{\tau} = N\bar{y} = 60{,}000 \times \frac{1073}{1203} =$ 53,516.2 . **[8]**

(e) [12] $\hat{p} = \frac{\hat{\tau}}{N_1} = \frac{53,510}{75,000} = 0.714$, **[5]**

$$\hat{Var}(\hat{p}_1) = \hat{Var}(\frac{\hat{\tau}}{N_1}) = \hat{Var}(\frac{N}{N_1}\bar{y}) = Var\left(\frac{60,000}{75,000}\bar{y}\right) = Var(0.8\bar{y}) = (0.8)^2 Var(\bar{y})$$

$$= 0.64\frac{S^2}{n}\frac{N-n}{N} = 0.64\frac{0.278}{1203}\frac{12000-1203}{12000} = 1.3307\times10^{-4}, \textbf{ [5]}$$

so that $B_1 = 2\sqrt{\hat{Var}(\hat{p}_1)} = 0.023$, **[2]**

[20] 3) Central Region public school board conducted a preliminary study about students graduating with A average. From each of the three areas in the region a few schools were sampled at random. The number of students expecting to graduate with A average was counted in each school. The results are given in the following table:

Area	Number of schools	Number of schools sampled	Sample results $\bar{y}_i$	S_i^2
1	45	9	82	30
2	36	7	80	20
3	20	4	56	30

(a) [7] What type of sample design was used here? Assuming that all schools are listed first by the area, and then by their name inside the area, explain in some detail how the sample can be selected using the table of random numbers listed below. What schools from area 3 would be selected using your method? Explain.

92325 19474 23632 27889 47914 02584 37680 20801 72152 39339 34806 08903 25570
31624 76384 17403 53363 44167 64486 64758 75366 76554 31601 12614 33072 60332
01624 76384 97403 53363 44167 64486 64758 75366 76554 31601 12614 33072 19474
23632 27889 47914 02584 37680 20801 72152 39339 34806 08930 25570 33120 45732

(b) [9] Estimate the average number of students per school in the region expecting to graduate with A average, and place a bound on the error of the estimation.

(c) [4] Can you estimate from the sample the proportion of students expecting to graduate with A average in the region? What information should be included in the sample to be able to estimate that proportion? What population information may already be available in the school board that may be combined with the sample to estimate the proportion?

(see next page)

Solutions:

(a) Sample design: stratified random sampling. Sampling units: schools. Strata: three areas. Population size N = 45+36+20=101. **[3]**

$N_1 = 45$, $N_2 = 36$, $N_3 = 20$. For sampling from each stratum we may use two digits from the table, assign properly to stratum elements, and read from the table until the required sample size is obtained. We may be more efficient if we assign more than one group of two digits to sampling units, such as

Area 1 schools	1	2	3	…	45	ignore
Digits	01, 51	02, 52	03, 53	…	45, 95	46,47, …, 50, 96, 97, …, 99, 00

Area 2 schools	1	2	3	…	36	ignore
Digits	01, 51	02, 52	03, 53	…	36, 86	37, 38, …, 50, 87, 88, …, 99, 00

Area 3 schools	1	2	3	…	20	nothing to ignore
Digits	01	02	03		20	
	21	22	23		40	
	41	42	43		60	
	61	62	63		80	
	81	82	83		00	

If we use this method for Area 3, and start reading from the first row, we get

Digits	92	32	51	94	74	23
Sampling units	12	12	11	14	14	03

After ignoring repeated digits (we use sampling without repetition), sampled schools are 3, 11, 12, 14. (any correct method is acceptable) **[4]**

(b) $\hat{\mu} = \frac{1}{N}\sum N_i \bar{y}_i = (45\times 82 + 36\times 80 + 20\times 56)/101 = 76.14$, **[4]**

$\hat{Var}(\hat{\mu}) = \sum \left(\frac{N_i}{N}\right)^2 \frac{N_i - n_i}{N_i}\frac{S_i^2}{n_i} = \left(\frac{45}{101}\right)^2 \frac{45-9}{45}\frac{30}{9} + \left(\frac{36}{101}\right)^2 \frac{36-7}{36}\frac{20}{7} + \left(\frac{20}{101}\right)^2 \frac{20-4}{20}\frac{30}{4} = 1.01503$, **[3]**

$B_\mu = 2\sqrt{1.01503} = 2.015$. **[2]**

(b) The proportion of students expected to graduate with A cannot be estimated from the sample because the information on the total number of student in the final grade (that are expected to graduate) is missing from the data. The total number of students that are expected to graduate should be available in the board, and it can be used to estimate the proportion. **[4]**