Last name: First name: Student #:

STA 304H1 F/1003H F SUMMER 2010, First Test, May 27 (20%)
Duration: 60 min. Allowed: hand-calculator, aid-sheet, one side, with theoretical formulas and definitions only.

[32] 1) The faculty senate at the University of Toronto wanted to know what proportion of students thought a foreign language should be required for everyone. With a help of the Department of Statistics a simple random sample of 500 students was selected from all students enrolled in statistical courses. A survey form was sent by e-mail to these 500 students. Answer in short the following questions:
(a) What is the population of interest to the faculty senate?
(b) What is the sampling frame?
(c) Describe the variable of interest. What type of the variable is the variable of interest?
(d) Discuss in short to what extend each of the three types of bias would be likely to occur in this survey: (i) inadequate frame, (ii) selection bias, (iii) nonresponse bias. Which of these three types of bias do you think would be the most serious in this study? Explain.
(e) Do you expect that the obtained estimate would overestimate, or underestimate the parameter of interest? Explain.

Solution:
(a) All students at the University, but it could be more restrictive, say, only undergraduate students. **[5]**
(b) All students enrolled in statistical courses. **[5]**
(c) Variable of interest: student's opinion whether a foreign language should be required for everyone, with values "Yes", "No". [**4**] It is a qualitative variable. **[2]**
(d) (i) The frame obviously does not cover the population, but the problem is whether it represent the population properly. The frame may not be representative for the population because students from some departments (e.g. languages and arts) are usually not interested in statistics and then will not be properly represented in the sample. **[4]**
(ii) There is no selection bias from the frame (SRS), but because of (i), it could be biased in comparison to the population. **[3]**
(iii) It is likely that not many students will respond due to lack of interest, or time. **[2]**
[other answers are acceptable, if they are justified]

It seems that the nonrespose bias will be the most serious, because it is likely that most of the students that are in favor of foreign language requirement will respond. Selection bias and inadequate frame are less significant, because most of students are enrolled in some statistical course. **[3]**

(e) An overestimation of the proportion in favor of foreign language may be expected, due to nonresponse bias (see above in (d)). **[5]**

[37] 2) Distribution of family sizes in a recent census in Toronto was as follows:

Number of persons	1	2	3	4	5	6	7	Total
Proportion of families, %	25	32	17	16	7	2	1	100%

(a) What is the population in this question? What is the variable? What other variables might be of interest in this population? Name a few.
(b) Calculate the population mean and standard deviation from the distribution in (a).

Using a list of households, a random sample of 400 families from the population was selected and the following result was obtained:

Number of persons	1	2	3	4	5	6	Total
Number of families	105	140	75	62	14	4	400

(c) Estimate the population mean and calculate the exact error of estimation (use a result from (b)).
(d) Calculate a bound on the error of estimation in (c) using a result from (b) (don't use the sample). Does the error of estimation from (c) indicate something about the sample and the census? Discuss.
(e) If you were to estimate the average family size in Toronto with a bound on the error of estimation of at most 0.10, what would you suggest as the sample size? You have no other information except that the family size is at most 7 (a few exceptions may be ignored), and, obviously, at least 1.
(f) Can you estimate the total number of people living in Toronto using only the sample given above? Why, or why not? Make a reasonable guess about the missing information and then estimate that number.

Solution:

(a) Population: All families in Toronto in the recent census. **[2]**
Variable: Family size. **[2]**
Other variables: Number of children, family income, family type, accommodation type, living in a house or apartment, ... **[2]**

(b) $\mu_y = \sum y_i p(y_i)$ = 1 x 0.25 + 2 x 0.32 + 3 x 0.17 + ... = 2.58, **[3]**

$\sigma_y = \sqrt{\sum (y_i - \mu_y)^2 p(y_i)} = \sqrt{\sum {y_i}^2 p(y_i) - {\mu_y}^2}$ = ... = 1.387 (σ_y^2 = 1.9236). **[3]**

(c) $\hat{\mu}_y = \bar{y} =$ (1 x 105 + 2 x 140 + ...)/400 = 2.38. **[3]**

Error of estimation = $|\hat{\mu}_y - \mu|$ = |2.38 - 2.58| = 0.20. **[3]**

(d) Error bound: Calculate first $SD(\hat{\mu}_y) = \sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}} = \frac{1.387}{\sqrt{400}}$ = 0.0693. **[3]**

(finite population correction is $\frac{N-n}{N-1} \approx 1$ because N is large)

Then B_y = 2 x 0.0683 = 0.1386. **[2]**

Error of estimation (0.20) is greater than B_y (0.1386). This may indicate that the sample is not properly selected, i.e. it is not an SRS. Or, it may be that the structure of families changed dramatically from the recent census, which is less likely. Obviously, it also may happen by chance (error of the first kind) **[2]**

(e) Estimate σ by $\frac{Range}{4} = \frac{7-1}{4} = 1.5$ [3] (not far from the value in (b)), and then

$$n \approx \frac{\sigma^2}{D} = \frac{(1.5)^2}{(0.10/2)^2} = \left(\frac{1.5}{0.05}\right)^2 = 900 .\ [3]$$

(f) It is not possible. The sample units are households (families), from which the sample of size n = 400 was selected. The total number of people living in Toronto is then the total of the population, and then $\hat{\tau} = N\hat{\mu}$, $\hat{\mu} = 2.38$, but information on *N* is not given. [3]

If we guess that N = 1,000,000 (or something about that value is reasonable), then $\hat{\tau} = 1,000,000 \times 2.38 = 2,380,000$. [3]

[31] 3) There are 850 students in an introductory statistical course at U of T, and their names are listed in a file with identification numbers 1, 2, ... , 850. At the beginning of the course, the instructor wants to conduct a short test to estimate how the class is prepared for university.

(a) Use the table of random numbers given below to select an SRS (without replacement) of 5 students from the first 70 students in the file. Select the sample and explain your method.

(b) Select also another SRS of 5 students, using idents from 71 to 500. Select the sample and explain your method.

(c) If you combine these two samples in one sample of size 10, would it be an SRS from the population? Explain why, or why not.

(d) Even if the sample in (a) may not be, strictly speaking, an SRS from the population, can it be, in a way, representative of the population (ignore small sample size)? You may consider some additional assumptions about the file. Explain.

(e) If an SRS of size 50 was selected, and found that 40 students didn't have any previous statistical background, estimate the total number of students in the class without any previous background in statistics, and calculate the confidence interval for the estimate.

Table of random numbers:
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

Solution:

(a) N = 70. Use two digits from the table of random numbers, e.g. pairs 92 **32 51** 94 74 **23 63 22** 78 89. [3] Sample: 22, 23, 32, 51, 63. **[3]**

(b) Use three digits, e.g. from the second row: **316 247** 638 **417 403 533** 634 416. Ignore digits below 071 (i.e. 000, 001, ... , 070) and above 500. **[3]**
Sample: 247, 316, 403, 417, 533 (or appropriate to the part of the table used). [3]

(c) The sample is not an SRS from the whole population [3], because elements 1, 2, …, 70 have higher chance of being selected, and elements 501, 602, …, 850 have no chance of being selected. **[4]**

(d) If the students in the file are in a random order, noncorrelated with their statistical background (e.g., all ordered alphabetically), then a sample from the first 70 students on the list will still be representative of the class. [3] If the students are ordered, e.g. first by departments, the sample may not be representative of the class because all sampled may com from the same department. **[3]**

(e) $\hat{p} = \bar{y}$ = 40/50 = 0.8 = 80% students without stat. background.
$\hat{\tau} = N\hat{p}$ = 850x0.8=680. **[3]** $B_\tau = N \times B_p = 850 \times 2 \times \sqrt{\frac{850-50}{850}\frac{0.8\times 0.2}{49}}$ =94.24, CI for the total is then 680±94.24 = [585.76, 774.24]. **[3]**