Last name, first name: . Student #:

STA 304H1 S WINTER 2013, First Test, February 15 (20%)
Duration: 55 min. Allowed: nonprogrammable hand-calculator, aid-sheet, one page two sided, or two pages one-sided, with theoretical formulas only, as posted on the web-site; the test contains 6 pages, please check. You may use any back side, with clear indication of Question part.

[35] Question 1) An experimenter wants to estimate the average water consumption per family in a large city.
(a) Discuss the relative merits of choosing the following cases as sampling units.
(i) Individual families. What information is needed? What would you use as a frame?
(ii) Dwelling units (single-family houses, apartment buildings, and so on). What information is needed? What kind of frame may be available?
(iii) City blocks. What information is needed? What kind of frame may be available? Where may you look for a frame? **(continued)**

Solutions:
(a) **[5]** (i) This would be best to use, as the variable of interest is a characteristic of a family. The problem is that we would need a list of all families in the city, which would be hard to find directly. The telephone directory would be a frame that would cover most of families. Single families and younger people, such as students, may be missing from the frame, as they nowadays use cell-phones, not listed in the directory. Other families, such as with nonlisted phones, or families that moved, may not be covered.
[5] (ii) Dwelling units would make more convenient sampling units, because there are much less changes of them over time. It would be necessary to have information on the number of families for every dwelling. Some of problems might be that some of dwelling units are unoccupied for some period; they could be quite different in size, such as apartment buildings with many families, which might require two stage sampling, which also would give a higher chance of selection to small dwellings. List of dwellings could be possible to find in the City Office.
[5] (iii) City blocks are easily accessible and there are not many of them. Detailed maps exist from which city blocks could be sampled. But, we would need to know how many families live in every block and how to measure water consumption, in every block. This may not be possible even for the City Water Supplier. Again, we may try then to use two stage sampling of families, which also requires sme frame inside every block.

Question 1, continued

(b) In which of the above cases in (a), an appropriate frame would be easiest to create?
(c) In each of the above cases in (a), explain in short which sampling design would be appropriate. Don't forget, elements of target population are individual families.
(d) From a previous study, the standard deviation of water consumption per day per family is estimated to be $\sigma = 0.2$ m^3. Find sample size of an SRS of families required to estimate the average consumption per family with an error bound of 0.04 m^3.

Solutions:

(b) **[3]** Obviously, the easiest frame can be obtained in (a) (iii), but that one would not be appropriate for the goal of the study. **[3]** Possibly the frame in case (a) (ii) would me the most appropriate, and would be possible to create from the information available in the city. ///**[6]**

(c) **[2]** In case (a)(i), an SRS would be appropriate, as the sampling units would be the families, which would give them all the same chance of being selected. **[2]** In case (a)(ii), a two-stage cluster sampling would be appropriate, with clusters of unequal sizes, often with just one family, or much bigger, with apartment buildings. **[3]** In case (a)(iii), the blocks may even be used as strata, combined with two, or three-stage cluster sampling, depending on block sizes. ///**[7]**

(d) **[2]** The city is large, so that we can use an approximate formula for calculating the sample size in an SRS, $n = \sigma^2/D$, where $D = (B/2)^2$.

[5] $D = (0.04/2)^2 = 0.0004$, $n = (0.2)^2/0.0004 = 0.04/0.0004 = 100$. ///**[7]**

[35] Question 2) State park officials were interested in the proportion of campers who consider their camping space adequate in a particular campground. The camping site has 500 spaces available, enumerated from 1 to 500. At a certain day, N = 350 places were occupied. The official decided to take an SRS of occupied camping sites from the park on that day (they count the number of visitors, but they don't know exactly what places are occupied).

(a) What is the target population?
(b) What will you use as a frame? What are the sampling units?
(c) Explain in detail (but don't complicate!) how to select an SRS of n = 30 occupied camping places from the park, using the frame, and the following portion of TRN

17403 53363 44167 64486 64758 75366 76554 31601 12614 33072 60332 92325 19474
23632 27889 97403 02584 37680 20801 72152 39339 34086 43218 15263 31624 76384
01624 76384 47914 53363 44167 64486 64758 75366 27889 31601 12614 33072 19474
23632 76554 47914 02584 37680 20801 72152 39339 34806 08930 25570 33120 45732

Show your procedure and the result for the first five sampling units. **(continued)**

Solutions:

(a) **[5]** Target population: the campers that are potential visitors to the campground.

(b) **[5]** A map or a list of all camping places in the park. Sampling units are camping places, as they are easily accessible from the frame.

(c) **[3]** As there are N = 500 places on the frame, we can use 3 digit numbers for selection and assign camping places (CP) to 3 digit numbers by

1 2 3 …. 500 ignore
001 002 003 …. 500 501 502 … 999 000

We also may improve the procedure by assigning two groups of 3 digits to a sampling unit, such as 001, 501 to 1,…, 499, 999 to 499, and 500, 000 to 500. The first method will be utilized below.

[3] Then, just by reading the table, e.g., using the first line, and first three digits out of five, we get

174, ~~533~~, 441, ~~644~~, ~~647~~, ~~753~~, ~~765~~, 316, 126, 330, …

[3] We proceed until we get 30 elements from the list, **but we also prepare several more cases, if some of selected camping places are unoccupied**. For every one of the unoccupied places we have to add one more element to our list, until we obtain 30 occupied places. ///**[9]**

Question 2, continued

(d) After consulting the sampled camping parties (one party = one camping site), it was found that 25 of them think the camping space is adequate. Estimate the total number of camping parties that are satisfied with the camping site, and place a bound on the error of estimation.

(e) (i) If the sampling and estimation is repeated on another day, what difference would it make? Explain. (ii) Do you think the results would be similar? Would they differ just by chance, or there would be another reason they may be different? Explain in short.

Solutions:

(d) We are estimating the proportion first, and then total.

$\hat{p} = \frac{a}{n} = \frac{25}{30} = 0.8333 = 83.33\%$ of satisfied camping parties **[2]**

$\hat{\tau} = N\hat{p} = 350 \times \frac{25}{30} = 291.67$ **[3]**

$\hat{\sigma}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{30-1} \times \frac{350-30}{350}} = \sqrt{\frac{0.8333 \times 0.1667}{30-1} \times \frac{350-30}{350}} = 0.06618$ **[2]**

$\hat{\sigma}(\hat{\tau}) = N\hat{\sigma}(\hat{p}) = 350 \times 0.06618 = 23.16$

$B_{\tau} = 2 \times \hat{\sigma}(\hat{\tau}) = 2 \times 23.16 = 46.32$ **[2]** ///**[9]**

(e) (i) **[3]** The main difference would be in the number of camping parties, that is, in the sampling population.

(ii) **[2]** They can be expected to be similar, up to random error, unless the conditions of the camping site change, when they may be significantly different, **[2]** e.g., when the weather is bad, when it started raining, which might make camping site less likable or convenient. ///**[7]**

[30] Question 3) A sociological study conducted in a small town called for the estimation of the proportion of households that contain at least one member over 65 years of age. The town has 621 households according to the most recent city directory. An SRS of size 60 households was selected from the directory. A summary of the results after the completion of the field work is given in the following table.

Number of members over 65 years of age	0	1	2	3	Total
Number of households	49	6	4	1	60
Total number of members	175	26	17	5	223

(a) (i) Estimate the true percentage of the households that contain at least one member over 65 years of age, and (ii) place a bound on the error of estimation.
(b) (i) Estimate the total number of members in the community. (ii) Can you place a bound on the error of estimation? Why, or why not? Explain. If you can, calculate it.
(continued)

Solutions:
(a) **[10]** (i) There are a = 11 households with at least one member over 65. **[1]**

$\hat{p} = \frac{a}{n} = \frac{11}{60}$ = 0.1833 = 13.33% of households with at least one member over 65 **[4]**

(ii) $\hat{\sigma}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{60-1} \times \frac{621-60}{621}} = \sqrt{\frac{0.1833 \times 0.8167}{60-1} \times \frac{621-60}{621}}$ = 0.04788 **[3]**

$B_p = 2 \times \hat{\sigma}(\hat{p}) =$ 2 x 0.04788 = 0.09575 = 9.57% **[2]**

(b) **[7]** (i) $\hat{\tau} = N\hat{\mu} = N\overline{y} = 621 \times \frac{223}{60} =$ 2308.05 **[4]**

[3] (ii) To place a bound on the error of $\hat{\tau}$ you need to calculate S^2 from the data, but it is not possible, because we don't have information on individual family sizes for all families. The sizes for families without seniors are not given; we only have the total for this group of families. **[only the correct answer is accepted here, no some ideas and unclear answers]**

Question 3, continued
(c) (i) Estimate the total number of members over 65 years of age in the community. (ii) Can you place a bound on the error of estimation? Why, or why not? Explain. If you can, calculate it.
(d) Using the results from (a)-(c), give a reasonable estimate of the percentage of the members **in the community** over 65 years of age. (don't think about theory)

Solutions:
(c) **[8]** (i) Use the average number of seniors per family.

$$\hat{\tau}_1 = N\hat{\mu}_1 = N\bar{y}_1 = 621\times\frac{49\times 0+1\times 6+2\times 4+3\times 1}{60} = 621\times\frac{17}{60} = 621\times 0.2833 = 175.95 \textbf{ [3]}$$

(ii) The data on the number of seniors per family is available for all families (for families without seniors, this number = 0), so that we can calculate S_1^2 and then calculate error bound on $\hat{\tau}_1$.

$$S_1^2 = \frac{1}{60-1}\left[49\times(0-\bar{y}_1)^2 + 6\times(1-\bar{y}_1)^2 + 4\times(2-\bar{y}_1)^2 + 1\times(3-\bar{y}_1)^2\right] = \frac{26.1833}{59} = 0.44375,$$

$$\hat{\sigma}(\hat{\tau}) = N\sqrt{\frac{N-n}{N}\times\frac{S^2}{n}} = 621\sqrt{\frac{621-60}{621}\times\frac{0.44375}{60}} = 50.7599$$

$$B_\tau = 2\times\hat{\sigma}(\hat{\tau}) = 2\times 50.7599 = 101.52. \quad \textbf{[5]}$$

(d) **[5]** Using the result from (c) (i), and from (b) (i), $\hat{p} = \frac{175.95}{2308.05} = 0.076233 = 7.62\%$ proportion of seniors over 65 years of age.