B. Exercises

for January 24, 2012

You should be prepared to present the exercises and at least an attempt to solve them. A complete solution is not required but you should at least have read, understood and attempted to solve them before going to the lecture.

1. For which of the following sampling problems can it be reasonable to assume a finite or an infinite population. (i.e. to estimate population parameters or the parameters of the latent model). Motivate your answer with about 10 lines per question!

a) A higher executive wants to study how the employees feel in a company with 1575 employees. He selects 55 workers out of 923, 35 white collar officials out of 502 and 10 higher officials out of 150. These are asked to fill in a questionnaire about the work environment and perceived health and are offered a medical check up.

b) A municipality asks 477 parents on how satisfied they are with the childcare in the municipality (which has 2 403 preschool children between 1 and 6 years). Only one parent of each child is asked (but if they have more children different parents can be selected for different children)

c) The Swedish Road Authority wants to know how the speed associated with the risk of accidents. They select 1000 cars passing a speed camera (out of 15 233). The owners get a questionnaire which among other things, asks about the number of incidents in recent years. They perform a loglinear regression between the number of incidents and the observed velocity.

d) In a study of the risks at childbirth of complications from computer work for the birth children 1,000 women who work at Social Insurance Office were selected. They were asked if they had had spontaneous abortions or given birth during the past year. They who had given birth were asked also if they had had any complications. For the selected women the computer logs were checked for how long they had actually worked at the computer during pregnancy.

2. A person has made an enquiry to 30 randomly selected families from the 231 households in a certain city area with large villas about the number of cars in the households. The following were his results

Number of cars	0	1	2	3	5	11
Number of households	7	11	7	3	1	1

a) Estimate the mean number of cars per household in the area. Estimate its standard error.

b) Compare the sensitivity of the estimate by making the same computations if the answer 11 is wrong (it is not clearly written on the questionnaire) and the true interpretation of the blurred value should be 1) (Some persons write 1 as a V turned upside down).

c) Study the intervals estimate +/- 2*standard deviation in both cases. Compare! Discuss the normality assumption and robustness. What would you advice the person to do when he presents the result of the study.

3. Consider a population with one hundred individuals. You should draw a sample of individuals and ask if they have a permanent job. From the sample you should estimate the proportion of permanently employed. Discuss the advantages and disadvantages of systematic sampling compared to simple random sampling. The sample size should be around 15.

The population from which you shall draw is the following (ordered after age)
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 1 0 1 1 1 1 1 1 0 1 0 0 1 0
0 0 1 1 1 1 1 0 1 1 1 0 0 1 0 1 0 1 0 1 1 1 0 0 0 1 1 1 0 0 1 1 1 0 0 0 1 0 0 0

Compute the variances of the two methods. Either by computing the correct value (You have the whole population) or by simulation by doing at least 5 independent samples from each method.

4. Choose a positive number!

a) Create a population by generating 100 random numbers uniformly distributed between o and the chosen number. Forget the chosen number!

b) Select a sample with the size 10 from the created population with SRS without replacement.

c) Estimate the mean value in the population (design-based) and estimate its standard error.

d) Estimate the chosen random number (which you have forgotten) using that you know that the observations are uniformly distributed. (The bias-corrected ML-estimate is (1 + 1/n) times the largest value in the sample. Estimate the mean value in the population (model-based).

e) Estimate the standard error of the estimate in d).
(This is not easy. But you may use that the variance of a uniform variable on (0,1) is 1/12 and that the maximum if n uniform random variables has the variance $n/((n+1)^2(n+2))$)

5. A person has done the following study. He has chosen 2 municipalities in Skåne (Scania), randomly with replacement and with inclusion probabilities proportional to the population. After that he selected 50 respondents in each of the two municipalities (100 different persons if the municipality was chosen twice). (This is not an easy exercise. It illustrates that it is not always easy to compute inclusion probabilities).

a) Compute the probability that a specific person in Höör is included in the sample. (You have to look up the sizes of the municipalities of Skåne).

b) Compute the probability that a specific person in Lund and a specific person in Höör both are included.

c) Suppose that he instead had merged the two selected municipalities and selected 100 with SRS from the combined population. Compute the same probabilities with this sampling plan.

6. A harbour was during the twelve months 2010 visited by 12, 7, 13, 15, 22, 21, 19, 23, 32, 29, 27, 32 vessels, respectively. Every month two randomly chosen vessels (the captain) is asked about how satisfied they were with the facilities and the service in the harbour (graded on a scale from 1 very dissatisfied to 10 very satisfied). Results:

Data

Month	J	F	M	A	M	J	J	A	S	O	N	D
Boat 1	7,	8,	8,	4,	3,	7,	5,	4,	7,	3,	9,	2
Boat 2	9,	7,	7,	5,	6,	4,	8,	3,	6,	6,	4,	5

a) Assume first (wrongly) that all 24 vessels is a simple random sample from all the 252 vessels. Estimate the average (which would have been obtained if all vessels had been sampled) and calculate the standard error of the estimate

b) Estimate the mean for each month with the standard error.

c) Weight these estimates together with weights proportional to the number of vessels each month. Compare with the result in a). Which one is best?

d) Compute the variance of the estimate from c. Compare with the result from a). Which one is smallest?

e) Discuss the use of a finite population here.

7. Suppose that you have made a political poll based on a simple random sample of 3 000 individuals. (What they should have voted for if there had been a parliament election today but also their opinions about some other issues like “job deduction” (jobbavdrag), “deduction for household-close-services” (RUT), Swedish wolves, “defence tax” (värnskatt), immigration issues and the asylum of political refugees, privatisation of schools, pharmacies and hospitals). You can from different registers complete the frame so that you know each individuals age, gender, education, assessed income, zip code, type of home (owned villa, rented villa, owned apartment, rented apartment and farmstead). Do you believe that these registers would help in the estimation? How? (Mainly a discussion point)