A. Take home assignments, Sampling and estimation, winter semester 2012

Can be handed in personally to Daniel Thorburn at the lectures, by post or e-mail Daniel.Thorburn@stat.su.se , out in the departments mailbox (outside the elevators) or in his department post-box or handed to Nicklas Pettersson. The assignments must have reached DT or NP before 17.00.

The solutions must be your own. You are allowed to discuss the problems with each other on a superficial level but two solutions that resemble each other too much will get less credit (e.g. the same calculation errors or paragraphs with the same wordings). Note also that these assignments will influence the marks on the course. Some of them are difficult in order to be able to differentiate between you. Do not expect to be able to solve everything perfectly.

For Monday, January 30

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

a) An auditor wants to find out the number of errors in the accounts of a savings bank. He selects 500 accounts out of 15 373 and writes to the account owner and asks whether the amount in the books are the correct value of their deposits in the bank at the end of the year. 479 of the account owners agree that this is the correct deposit. The remaining 21 are more closely examined and it turns out that in 12 cases the figures from the bank were correct but in 9 cases transactions were booked at the wrong year.

b) In a study of whether dyslexia increases in a Sweden 200 schools are asked about the number of pupils in the fifth grade who have a diagnosis of dyslexia (and the total number of pupils in that grade). In the study the schools are also asked about the measures they have taken to ensure that the pupils get the support they are entitled to. This study is repeated the following year.

c) A shop owner wants to know how their customers are treated in the shop and their opinion of the products for sale in the shop. A market institute hires two interviewers who stand outside the shop and interviews a random sample of the customers leaving the shop.

d) Statistics Sweden tries to estimate the gross national product of Sweden. A sample of enterprises are, in some detail, asked about their sales and their costs during January (and also about their production and changes in inventories).

e) 3245 randomly selected Swedes are asked about their attitude in a series of political issues. One wanted to know including their attitudes to the European Union, to the Euro, their political sympathies and their attitudes to nuclear power.

2. At a university a simple random sample of students are after the first test asked about how long much time they have allocated to the studies per week (non-anonymously and on the

web). When the sample is analysed it is divided into four groups. Did not partake in the test (N=48, n=12), did not pass (N=27, n=8), pass (N=52, n=14) and high pass (N=25, n=6).

Data not partake	8, 8, 24, 22, 7, 0, 0, 12, 7, 19, 2, 15
Not pass	27, 12, 8, 5, 17, 4, 25, 2
Pass	27, 21, 19, 15, 12, 28, 32, 20, 20, 20, 11, 17, 20, 14
High Pass	25, 20, 25, 25, 30, 10

Estimate the average number of hours allocated to the studies in the four groups, estimate the variances of the estimates and give uncertainty intervals.

Here you should view the problem as a finite sample problem. Discuss whether this is a reasonable assumption.

3. Today a common way of doing surveys is by using web-panels. A common way of doing such a survey is to first select a large random sample from the (Swedish) population. All of them are asked if they agree to be part of a web-panel and also asked to give their e-mail addresses. They are also asked a lot of background questions. This panel is then used as a frame in many studies in the future.

The real survey is done by selecting a random sample from the panel created above, sending out a questionnaire on the web. This gives a fast and cheap study since the answers go directly into the computer and are already checked and edited. You can get make a study in one week with a sample of 10000. The sample is in principle taken with probabilities inversely proportional to the tendency to agree to be in the panel in the first study. E.g. if 30 % of the women said yes and 15 % of the men, this is compensated by selecting men with twice the probability in the web survey. The web sample will thus have the correct proportions in all known background variables (e.g. age sex, region but also income, education, social group and political opinion (vote in last election) ...)

Discuss whether this is a recommendable procedure. Write an essay with a length corresponding to about one page in Word. Discussing the pros and cons?

4. Spooose that you are working as a consulting statistician. You are asked to construct an estimating plan for the following study. The client had had access to three lists on traffic accidents the road: Police reported accidents, patients hospitalized for traffic injuries and claims reported to insurance companies (with N1, N2, N3 cases on the three lists). One accident may be on none, one, two ort three of the lists. On the second and third list several cases may correspond to the same accident. The client had selected n1, n2 respectively n3 cases with SRS from the three lists and eliminated duplicates. The intended sampling frame was the merged list.

An incident that appears on multiple lists has thus a greater probability of selection than one who appears only once. The number of times it is recorded probably has some connection with the accident character. You realize that in order to obtain unbiased estimates you must

take the first and second order inclusion probabilities into account (and that they are functions of how many times the accident appears on the three lists).

Your task is to write a short PM or report to your principal pointing out the problems with varying inclusion probabilities, computing or suggesting ideas on how the inclusion probabilities can be computed, suggesting what questions to ask in order to be able to compute the probabilities.

To pass it is not required that you resolve all the parts - but you should justify this to the client and point out which problems remain.

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 0 1 0 0 1 1 0 1 1 1 0 1 0 1 1 1 1 1 1 0 1 0 1 0 0 1 0
0 0 1 1 1 1 1 1 0 1 1 1 0 0 1 0 1 0 1 0 1 1 0 1 1 1 0 0 0 1 1 1 0 0 1 1 1 0 0 0 1 0 1 0 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)

For January 30

8. Compute the first and second order inclusion probability of the following procedure. You select units from a list. The first unit is selected with probability 1/8. If the first unit is selected the second one is not selected, but if the first unit is not selected the second one is selected with probability 1/6. The selection is then continued in the same way. If a unit is selected the next one is not but if a unit is not selected the next one is included with probability 1/6. The reason for this procedure is that it is impractical to interview to interview two units which are close to each other.

9. In order to measure the knowledge of the pupils in geography, when they leave the ninth grade a pedagogue selected 10 Swedish schools with SRS. All pupils in the ninth grade in these schools were asked to complete a simple test with 120 multiple choice questions. These questions were designed to measure how well the pupils performed according to the goals of the Swedish School board. (The result 60 % (or 72 correct questions) was considered to correspond to the grade pass).

School	A	B	C	D	E	F	G	H	I	J
# pupils (in ninth grade)	180	121	45	52	73	141	170	16	106	45
Average number correct	75	80	99	87	73	78	92	112	89	102
Proportion pass %	59	58	87	85	47	53	70	100	91	95

There are altogether 102 000 pupils in Sweden and altogether 1200 schools.

a) Estimate how many correct answers an average Swedish pupil will get. (Help: put Y = the product of the first and second line and X = the first line)

b) Estimate the proportion of Swedish students who will get a pass.

c) Give the standard error of these estimates.

10. Someone has made a study of 100 supermarkets with SRS and he has asked about their turnover of dairy products, Y. The frame consists of 2300 supermarkets and it contains also the number of employees X_{1i} and the total turnover X_{2i} last year.

Data:
In the frame:
Total number of employees is 22 100 and the total turnover is 98 000 MSEK

In the sample:
Total sale of dairy products 320 MSEK, total number of employees last year 942, total turnover last year 3 900 MSEK.
Estimated covariance matrix

	Dairy products	Employees	Turnover
Dairy products	132 000	2 700	1 250 000
Employees	2 700	131	49 000
Turnover	1 250 000	49 100	22 100 000

a) Estimate the total amount of sold dairy products using a regression estimator with only total turnover as auxiliary variable.
b) Estimate the total amount of sold dairy products using a generalised regression estimator with both number of employees and turnover as auxiliaries
c) Estimate the standard errors of these estimates.