B. Exercises

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.

For February 20

17. A company delivering letters has 5 000 mailman tours over the whole country. The number of letterboxes is known on every tour (x_i; i=1, …, 5000). In total, there are 5 000 000 letterboxes in the whole country ($\Sigma_{i=1}^{5000}$ (x_i) = 5 000 000). One hundred mailmen are selected with probabilities proportional the number of boxes on their tours. (Pareto πps). Every mailman has to put a questionnaire in every tenth letterbox, where the person who usually empties the box is asked questions about the postal service and letter delivery. The questionnaire should than be posted to the mail company.

a) What is this sampling design called?

b) 10 000 questionnaires, with many questions, were handed out, We will only consider the question: ”What is your opinion on the proposal to move all letter-boxes in apartment houses to the entrances? Good, Bad, No opinion. (Mark the answer that best corresponds to your opinion)”. For every mailman tour the number of the three possible answers are counted (Data: (y_{iG}, y_{iB}, y_{iN}); such that $y_{iG}+y_{iB}+y_{iN} = z_i$ for i = 1, …, 5000) (partial non response is counted as no opinion). (The number of questionnaires handed out is also known, z_i if there had been non response the z_i would have been equal to $x_i/10$).

The company wants to estimate the proportion of households who consider this proposal as bad (out of all households in Sweden). Give and explain all formulas that you think they should use for estimation and their standard errors. State also all assumptions that you make)

c) Have you any comments on the sampling plan. Is it good in this case (for this question)? What do you think would happen if all answers were weighted equally?

18. A journalist wants to investigate the existence of “gender equality plans” in the companies owned by Swedish local authorities. (According to the Swedish law they need to have one). She selected 10 municipalities and two counties at random For the selected municipalities she found out the number of companies. She wrote to all those and asked for a copy of the plan. If she had not received a reply within a week, she called. When she published her article after a month, her data looked like this (they are not real)

Municipality	# companies	# received plans	# without plan	# no response	Sum no received plan
A	12	7	4	1	5
B	7	2	0	5	5
C	4	4	0	0	0
D	17	2	9	6	15
E	6	5	1	0	1
F	12	11	0	1	1
G	3	3	0	0	0
I	9	13	7	5	12
H	25	3	3	3	6
J	12	4	2	6	8
Subtotal	107	54	26	27	53
Sum of squares	1537	422	160	133	521
County					
Ö	6	6	0	0	0
Ä	9	5	2	2	4
Subtotal	15	11	2	2	4
Sum of Squares	117	61	4	4	16

a) Assume that the sample of municipalities is an SRS from Sweden's 290! Estimate the percentage of municipal companies which have been able to present a plan after a month!

b) Specify the uncertainty! (You may use that the product sum between the first and last column is 818)

c) Same questions for the county municipal companies (there are 20 counties)!

d) Estimate the percentage of all municipal corporations and specify the uncertainty, (County and municipal together)!

19. Every Monday a company receives a shipment with about 20 boxes containing a special detail (“shranks”) needed for the next week’s production. Every box contains 560 details. Every week two boxes are selected randomly and from each box 10 details, also randomly. These are checked carefully for defective. This procedure is prescribed in detail in the contract.

Data													
Week	1	2	3	4	5	6	7	8	9	10	11	12	13
No boxes	14	20	21	24	22	20	22	19	14	16	20	24	22
No defective													
Box 1	0	0	0	2	0	0	1	4	0	0	0	1	0
Box 2	3	0	1	0	1	0	0	3	1	0	0	0	0

a) Consider only the first week. Estimate the number and the proportion of defective in the first weeks delivery!
b) Give the standard error!
c) Now consider all 13 weeks. Estimate the proportion of defectives during the first quarter and state the standard error!

21. There is much debate in Sweden on the mathematical knowledge of nine grade students. Suppose that here is a internationally developed instrument (questionnaire) to measure the knowledge in mathematics on a scale between 0 and 100. The variance of the instrument (repeatability) is 25 and the variance between students over a population is said to be 400. The instrument is intended to be used in class of up to 50 persons in a class room surveilled by a specially trained person. The filling in the questionnaire takes about two hours (=three 40 minutes lectures). Someone has decided that this instrument should be used in Sweden and that the average knowledge should be estimated.

Your task is to construct a sampling plan. Your budget for data collecting is 2 MSEK and two months (counting costs but not time for training of personal). Suggest a plan. To do so you must determine some costs and variances which are not given here. Which costs and variances do you need other than those given? Suggest reasonable values for those and determine a reasonable sampling plan and give formulas for the estimation.

For February 27

22. A person has drawn a SRS-sample with n = 10 from a population with N = 50. Unfortunately one unit becomes partial nonresponse for one of the variables . The observations are

Y	112, 123, -, 122, 147, 110, 125, 123, 133, 111
X	35, 58, 27, 67, 119, 2, 55, 79, 107 , 23

a) The person does not worry about the non response, but makes the analysis as if he had a SRS with n=9, What will his estimate of the mean for Y_3 be and what will the variance of his estimate be?

b) Use the theory of linear regression and find the predictive distribution of Y for the two missing values.

c) The person reads about multiple imputation and he draws a new value from the predictive distribution of Y_3. He inserts this and makes the analysis as if he had had a complete data set. In his draw he got $Y^*_3 = 120$. What will his mean and variance be?

d) Repeat this nine more times. And gets nine more means and variances. Do this. Compute the mean of the ten means and the ten variances. Compute the variance of the ten means. Add the variance of the means and the mean of the variances. Compare these with the value that you found in a).

23. In a medical study a population of 12 patients are treated for a special type of heart failure, with a new method. All persons are before treatment classified into five categories after the severity. Half a year after treatment the patients are contacted once again and among other things the stability of the heart rhythm is measured. However, only 9 patients were reached.

Data:

Severity	Stability measurements
1	23, missing
2	15, 24, 12, missing
3	16, 21, 10
4	8, 13, missing
5	None

Your task is to do a multiple imputation.

a) Do a linear regression explaining the stability by the severity. Estimate the variance and covariance of the two parameter estimators, slope and intercept. (The variance around the line is from previous trials known to be 5)
b) Draw a pair of possible parameter values assuming that the true value is normal around the estimates with the variances/covariance from a). (Assume normality)
c) Draw three possible stability values for the missing assuming normality and that the parameters from b) are correct. Impute these values instead of the missing
d) Estimate the average stability and its variance for an infinite population assuming that the true values are the true values.
e) Repeat b-d four more times. You have now five mean value estimates and five variance estimates.
f) Find the imputation variance by finding the variance of the five repeated imputations.
g) Find the final estimate as the mean of the five repetitions and the final variance as the sum of the average variance from the repetitions plus the imputation variance.