

## Bonus assignments for MVE155 (Spring 2017)

There are four bonus assignments. The first is a modified version of an exercise in the course book, while the others are taken directly from the book. Correctly solving all assignments will lead to a maximum of 3 bonus points for the course exam.

The four assignments are associated with four different data sets. Links to these data sets may be found on the course home page. The purpose of the assignments is to make you go beyond purely analytical pen-and-paper problem solving. Therefore, some programming will be involved in each of the assignments. You are free to use the programming environment that you are most familiar with, but either R or Matlab will probably be the most convenient alternatives.

Please feel free to discuss ideas regarding the solution of the bonus assignments. However, both coding and report writing has to be done individually. If there are strong reasons to suspect that a student has copied either the code or some part of the report from another student, zero points will be awarded.

Teaching support for the bonus assignments will be provided by Sebastian (room L3098) at the following hours:

- Friday, Jan 27, 11:00-12:00.
- Friday, Feb 10, 11:00-12:00.
- Friday, Feb 24, 11:00-12:00.
- Friday, Mar 3, 11:00-12:00.

## Assignments

### Assignment 1 (survey sampling)

Read the description of the data set `families` given in problem 67 in chapter 7 of the course book (Rice, 3rd edition), but do NOT engage in solving (a-f) as stated there. Instead, do the following:

- a Take five different simple random samples, each consisting of 500 families. Estimate the following population parameters, calculate the estimated standard errors of these estimates and form 95% confidence intervals:
  - i The proportion of male-headed families.
  - ii The average number of persons per family.
  - iii The proportion of heads of households who received at least a Bachelor's degree.
- b Take 100 samples of size 400.
  - i For each sample, find the average education level of head of household.
  - ii Find the average and standard deviation of these 100 estimates and make a histogram of the estimates.
  - iii Superimpose a plot of a normal density with that mean and standard deviation of the histogram and comment on how well it appears to fit.
  - iv For each of the 100 samples, construct a 95% confidence interval for the population average education level.

- v Take 100 samples of size 100. Compare the averages, standard deviations, and histograms to those obtained for a sample of size 400 and explain how the theory of simple random sampling relates to the comparisons.
- c In the last question, you will analyse the family incomes.
- i Take simple random samples of size 400 from each of the four regions (North, East, South and West). Compare the incomes by region by making parallel boxplots.
  - ii Does stratification by family type help in estimating the average family income? Use a simple random sample of size 300 to estimate the average income and also the standard error of your estimate. Also, use the sample to form a 95% confidence interval. Next, allocate the 300 samples proportionally to the three family types and estimate the average income from the stratified sample. Estimate the standard error and form a 95% confidence interval. Finally, allocate the 300 samples to the three family types using optimal allocation and estimate the average income from the stratified sample. Estimate the standard error and form a 95% confidence interval. Compare the stratified results to the results obtained for the simple random sample. Hint: In order to perform the optimal allocation, you will need to estimate the variance in each strata first. For this assignment, it is OK to do this by using the sample variance while including all data for each respective strata.

### **Assignment 2 (parameter estimation, bootstrap)**

Solve problem 46 (a-g) in chapter 8 of the course book. The associated data set is `whales`.

### **Assignment 3 (sample comparison)**

Solve problem 51 (a-b) in chapter 11 of the course book. The associated data set is `ears`.

### **Assignment 4 (linear regression)**

Solve problem 37 in chapter 14 of the course book. The associated data set is `barium`.

### **Deadlines**

The deadlines for handing in the reports for the respective assignments are given below. In the interest of fairness, please respect them.

**Assignment 1:** Friday, Feb 3 at 23.59.

**Assignment 2:** Friday, Feb 17 at 23.59.

**Assignment 3:** Friday, Mar 3 at 23.59.

**Assignment 4:** Friday, Mar 10 at 23.59.

## **On the reports**

One report should be written for each assignment. Each report should be sent in the form of a PDF file to Sebastian (jobjorns@chalmers.se) before the deadline expires. Note the following points regarding the reports:

- Use English.
- Keep them concise but clear. You should demonstrate that you have understood the problem and the method of solution. Of course, the actual answers should also be given. There is no need to write a general introduction or a discussion section. However, if the problem statement in the course book explicitly asks you to discuss or explain something, then you should do this.
- Some results are best reported using plots.
- Include the code that was used to solve the assignment in an appendix.
- Do not forget to write your name and e-mail for further communication.