Computer Lab 5
Computational Statistics

Linköpings Universitet, IDA, Statistik

November 23, 2021

Course code and name:	732A90 Computational Statistics
Lab session:	29.11, 8-10
Submission deadline:	3.12, 23:59
Resubmission deadlines:	resubmission 1: 17.12, 23:59; resubmission 2 for labs 5-6: 21.1
Seminar:	Seminar 3 (first part) on 15.12
Teachers:	Maryna Prus, Filip Ekström, Joel Oskarsson, Martynas Lukosevicius, Shashi Nagarajan, Yifan Ding
Instructions:	This computer laboratory is a part of the examination Create a group report (in English) on the solutions to the lab as a .PDF file. All R codes should be included as an appendix into your report. In the report reference all consulted sources and disclose all collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to your teacher - see file ”lab groups” on lisam).

Exercises originally developed by Krzysztof Bartoszek

Question 1: Hypothesis testing

In 1970, the US Congress instituted a random selection process for the military draft. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. The first date drawn from the drum received draft number one, the second date drawn received draft number two, etc. Then, eligible men were drafted in the order given by the draft number of their birth date. In a truly random lottery there should be no relationship between the date and the draft number. Your task is to investigate whether or not the draft numbers were randomly selected. The draft numbers (`Y=Draft_No`) sorted by day of year (`X=Day_of_year`) are given in the file `lottery.xls`.

Compute an estimate $\hat{Y}$ of the expected response as a function of X by using a loess smoother (use `loess()`). To check whether the lottery is random, it is reasonable to use test statistics

$$T = \frac{\hat{Y}(X_b) - \hat{Y}(X_a)}{X_b - X_a}, \quad \text{where } X_b = \text{argmax}_X \hat{Y}(X), X_a = \text{argmmin}_X \hat{Y}(X).$$

If the expected value of T is significantly different from 0, then there should be a trend in the data and the lottery is not random. Estimate the distribution of T by using nonparametric bootstrap with $B = 2000$ (construct histogram). Using the obtained distribution (histogram) conclude whether the lottery is random or not.

Question 2: Bootstrap, jackknife and confidence intervals

The data you are going to continue analyzing is the database of home prices in Albuquerque, 1993. The variables present are `Price`; `SqFt`: the area of a house; `FEATS`: number of features such as dishwasher, refrigerator and so on; `Taxes`: annual taxes paid for the house. Explore the file `prices1.xls`.

1. Plot the histogram of `Price`. Does it remind any conventional distribution? Compute the mean price.

2. Estimate the distribution of the mean price of the house using (nonparametric) bootstrap. Determine the bootstrap bias-correction and the variance of the mean price. Compute a 95% confidence interval for the mean price using bootstrap percentile, bootstrap BCa, and first–order normal approximation
   (**Hint**: use `boot()`,`boot.ci()`,`plot.boot()`,`print.bootci()`)

3. Is the estimated mean located in all confidence intervals?

4. Estimate the variance of the mean price using the jackknife and compare it with the bootstrap estimate