Computer Lab 5
Computational Statistics

Linköpings Universitet, IDA, Statistik

2020/02/12

Kurskod och namn:	732A90 Computational Statistics
Datum:	2020/02/10–2020/02/28 (lab session 21 February 2020)
Delmomentsansvarig:	Krzysztof Bartoszek, Hao Chi Kiang
Instruktioner:	This computer laboratory is part of the examination for the Computational Statistics course Create a group report, (that is directly presentable, if you are a presenting group), on the solutions to the lab as a **.PDF** file. Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments. **All R code should be included as an appendix into your report.** A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes. 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 krzysztof.bartoszek@liu.se or hao.chi.kiang@liu.se), by **23:59 28 February 2020** at latest. Notice there is a final deadline of **23:59 5 April 2020** after which no submissions nor corrections will be considered and you will have to redo the missing labs next year. The seminar for this lab will take place **11 March 2020**. The report has to be written in English.

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`.

1. Make a scatterplot of Y versus X and conclude whether the lottery looks random.

2. Compute an estimate $\hat{Y}$ of the expected response as a function of X by using a loess smoother (use `loess()`), put the curve $\hat{Y}$ versus X in the previous graph and state again whether the lottery looks random.

3. To check whether the lottery is random, it is reasonable to use test statistics

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

   If this value is significantly greater than zero, then there should be a trend in the data and the lottery is not random. Estimate the distribution of T by using a non–parametric bootstrap with $B = 2000$ and comment whether the lottery is random or not. What is the p–value of the test?

4. Implement a function depending on *data* and B that tests the hypothesis
   H_0: Lottery is random
   versus
   H_1: Lottery is non–random
   by using a permutation test with statistics T. The function is to return the p–value of this test. Test this function on our data with $B = 2000$.

5. Make a crude estimate of the power of the test constructed in Step 4:

   (a) Generate (an obviously non–random) dataset with $n = 366$ observations by using same X as in the original data set and $Y(x) = \max(0, \min(\alpha x + \beta, 366))$, where $\alpha = 0.1$ and $\beta \sim \mathcal{N}(183, \text{sd} = 10)$.

   (b) Plug these data into the permutation test with $B = 200$ and note whether it was rejected.

   (c) Repeat Steps 5a–5b for $\alpha = 0.2, 0.3, \ldots, 10$.

   What can you say about the quality of your test statistics considering the value of the power?

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

4. Compare the confidence intervals obtained with respect to their length and the location of the estimated mean in these intervals.