

Computer Lab 5

Computational Statistics

Linköpings Universitet, IDA, Statistik

2022 XII 07

Kurskod och namn:	732A90 Computational Statistics
Datum:	2022 XII 05—2022 XII 13 (lab session 7 XII 2022)
Delmomentsansvarig:	Krzysztof Bartoszek, Bayu Brahmantio, Jaskirat Marar, Shashi Nagarajan
Instruktioner:	<p>This computer laboratory is part of the examination for the Computational Statistics course</p> <p>Create a group report, (that is directly presentable, if you are a presenting group), on the solutions to the lab as a .PDF file.</p> <p>Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.</p> <p>All R code should be included as an appendix into your report.</p> <p>A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.</p> <p>In the report reference ALL consulted sources and disclose ALL collaborations.</p> <p>The report should be handed in via LISAM</p> <p>(or alternatively in case of problems e-mailed to bayu_dot_brahmantio_at sign_liu_dot_se, jasma356_at sign_student_dot_liu_dot_se, shashi_dot_nagarajan_at sign_@liu_dot_se, or krzysztof_dot_bartoszek_at sign_liu_dot_se),</p> <p>by 23:59 13 December 2022 at latest.</p> <p>Notice there is a deadline for corrections 23:59 22 January 2023 and a final deadline of 23:59 12 February 2023 after which no submissions nor corrections will be considered and you will have to redo the missing labs next year.</p> <p>The seminar for this lab will take place 13 December 2022.</p> <p>The report has to be written in English.</p>

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 there can be doubts concerning the randomness of the selection of the draft numbers. The draft numbers ($Y=\text{Draft_No}$) sorted by day of year ($X=\text{Day_of_year}$) are given in the file `lottery.xls`. The data was originally published by the U.S. Government, and most conveniently made available online at http://jse.amstat.org/jse_data_archive.htm (see also Starr Norton (1997) Nonrandom Risk: The 1970 Draft Lottery, Journal of Statistics Education, 5:2, DOI: 10.1080/10691898.1997.11910534)

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

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

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 $\alpha = 0.2, 0.3, \dots, 10$.

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`. The source of the original is the Data and Story Library (<https://dasl.datadescription.com/>) and it can be recovered from (<https://web.archive.org/web/20151022095618/http://lib.stat.cmu.edu/DASL/Datafiles/homedat.html>).

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.