Statistics for Data Science (52015)

Research Project

Winter semester, 2022-2023

Submission guidelines

- The assignment is to be submitted via Moodle only.
- The deadline for submission is Monday, 13.02, 23:59. Projects submitted later than that without approval will be graded as zero.
- The total number of points is 105, but if you managed to score more than 100, your project score would be 100.
- You are allowed to submit in couples. Only one has to submit the project, but make sure you name your partner in the title of the PDF.
- The project can be programmed either in Python or in R. Not both. Choose your partner wisely.
- Your submission should include two files, either an R-Markdown file and a PDF of the output, or a Jupyter notebook file, and a PDF of the output.
- Each question is enumerated. Use those numbers when answering your questions.
- FIGURES: Your results should be self-explanatory and readable: each figure should have a unique informative title and each axis should have a meaningful label. Points will be reduced for ambiguous graphs.
- CODE: There are no special requirements or guidelines for your code, but you are expected to keep it organized and clear to understand (please add some comments when needed). Note that your code should run and output the same results as in your PDF. The seed (random generator) you use for your sampling should be the first four digits of (one of) your ID.

1 Estimation (35 points)

Adam and Eve had an argument about the average number of fruits each apple tree in heaven produces. The trees are assumed to be independent and identically distributed. Adam and Eve first tried to understand what parametric model (if any) to use:

a) For each value of the parameter, $p = 0.3, 0.5, 0.7$: Sample 50 i.i.d $Bin(100, p)$ random variables. This would represent 50 trees that Adam sampled, and counted the fruits in them. Build an approximate 95% confidence interval for the average, using:

   - Normal approximation
   - Parametric Bootstrap
   - Non-Parametric Bootstrap

   Compare the results. Does the parameter affect the length of the CI? Make sure to explain your calculations.

b) For $i = 1, ..., 1000$: Sample 50 i.i.d $Pois(50)$ and for each compute the average, such that at the end you have one thousand averages. Compute on those averages the coverage percentage of the Binomial confidence intervals with $p = 0.5$ you computed in (a). Explain your results.

c) Repeat (1.a) with 50 i.i.d samples from $Pois(50)$, and compare the confidence intervals of (1.a) and (1.c).

d) Suggest a distribution (known, or unknown) defined on the integers, with expectation 50, such that the coverage level of the confidence intervals from (1.c) will have a coverage level smaller than 95%. You may prove it analytically or show it numerically (repeat the method in (1.b)). How did you choose the distribution?

2 Hypotheses Testing (35 points)

Eve started suspecting that Adam is cheating when counting the apples on each tree, and he is falsely adding apples to each count. To verify it, she counted herself 50 independent trees (not Adam's trees).

a) Suggest a non-parametric test to Eve to check whether Adam is cheating, i.e., whether the two samples are from the same distribution, but not with the same expectation, with some confidence level $1 - \alpha$.

b) Test your suggestion using simulations (I.e., repeat the process multiple times), assuming the trees are distributed $Pois(50)$, and consider three cases: (i) Adam is not cheating, (ii) Adam is adding a constant number 5 to each count, (iii) Adam is sampling another independent $Pois(5)$ after he counts each tree, and he adds it to the count of the tree (new i.i.d $Pois(5)$ for each tree). **Clarification: You should sample both for Adam and Eve multiple times, and show that your proposed test is more likely to reject when Adam is cheating, compared to when he is not.**

c) Note that if Adam is cheating according to (iii), and if the apples have a $Pois(50)$ distribution, then Adam is actually sampling from $Pois(55)$ distribution. Suggest two parametric tests to determine whether Adam is cheating or not in this case (i.e., $H_0 : \lambda = 50, \quad H_1 : \lambda = 55$) with confidence level 95%. Show with simulations that you indeed obtained the desired confidence level.

d) Adam is willing to count more trees to show Eve he is not cheating. Under one of the models you suggested in (2.c), how many trees should Eve ask Adam to sample, such that the probability to detect if Adam is cheating, given that he is cheating, will be higher than 90%? (the power). You can either answer analytically, or use simulations. Make sure you report the minimal sample size.

3 Statistical Models (35 points)

Adam and Even collected many variables, trying to understand what affects the number of fruits of each tree. For example, the proximity to a water source, the amount of direct sun, the height of the tree, and so on. The trees are independent, but no longer identically distributed. They saved the results in the file "AppleTrees.csv". The first column is the number of apples, and all the other columns (24) are different covariates (their names do not matter).

a) Split your data randomly to train (80%) and test (20%) sets. Run a linear regression model with all the covariates (do not use transformations of any kind) using the training set. What is the SSE (sum of squared errors) of the training set? and of the test set?

b) Adam and Eve wish to find a subset of covariates that are the most important for the prediction task. Suggest two different methods, implement them, and report the subset with the corresponding SSE (for the training and the test). Why the SSE of the training set is larger than that of (3.a)?

c) Eve argued that the most important (for the prediction) covariates are also the most significant ones (i.e., with the smallest P.value). Respond to Eve, and use the data to support your response (feel free to use simulated data as well, if you wish).

d) Use the FDR method and the Bonferroni correction to multiple hypotheses and report what variables are rejected, with a confidence level of 0.95 (the FDR parameter is 0.05). Here you are requested to program the FDR on your own.

Good luck!