096202 - Introduction to Data Analysis

HW3 - Estimation and Hypothesis Testing

Guidelines

- Deadline is 02.01.20 at 23:55
- Submission in singles. Each student is expected to work on the notebook independently.
- Discussing the homework (but not working on them) with other students is allowed, but any student who does so must add a markdown cell at the top of the notebook explicitly naming the person with whom s/he discussed the homework.
- Submission will include a single Jupyter notebook file named '**HW3_ID.ipynb**'
- We will use the same dataset from HW2: 'HW2-Data.csv'
- Present all supporting calculations and analyses in **code cells**. Document your code with comments and clearly specify the final answers in **markdown cells**
- Any visualization **must include** axis labels.
- A particularly thoughtful and careful work may result in bonuses.

Dataset

Description

The data was collected in a study during the 2005-2006 school year from two public schools in Portugal. The dataset was built from two sources: school reports, based on paper sheets (i.e. grades and absences) and questionnaires. The dataset contains the following fields:

Column Name	Description
school	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	Student's sex (binary: 'F' - female or 'M' - male)
age	Student's age (numeric: from 15 to 22)
address	Student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status (binary: 'T' - living together or 'A' - living apart)
Medu	Mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Fedu	Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)
Mjob	other's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	Home to school travel time (numeric: 1 -<15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 ->1 hour)
studytime	Weekly study time (numeric: 1 -<2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 ->10 hours)
failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	Extra educational support (binary: yes or no)
famsup	Family educational support (binary: yes or no)
paid	Extra paid classes within the course subject (binary: yes or no)
activities	Extra-curricular activities (binary: yes or no)
nursery	Attended nursery school (binary: yes or no)
higher	Wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	With a romantic relationship (binary: yes or no)
famrel	Quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
goout	Going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	Workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	Current health status (numeric: from 1 - very bad to 5 - very good)
absences	Number of school absences (numeric: from 0 to 93)
G1	First period grade (numeric: from 0 to 20)
G2	Second period grade (numeric: from 0 to 20)
G3	Final grade (numeric: from 0 to 20, output target)
class	Name of course

Questions

1. It is known that 51% of the population in Portugal which was eligible for study in public schools in 2005 were females. Do the communities served by the two public schools in this dataset come from a simple random sample of the Portugal population? Check using the variable *sex*. (required significance level: 0.01)
   a) Clearly state the null hypothesis and the alternative hypothesis.
   b) What is the test statistic?
   c) Write code to test the hypothesis using simulations. Explain your code.
   d) What is your conclusion? Show both numerical result of the test and a plot demonstrating it.
2. Is the distribution of free time after school uniform? (required significance level: 0.05)
   a) Clearly state the null hypothesis and the alternative hypothesis.
   b) What is the test statistic?
   c) Write code to test the hypothesis using simulations. Explain your code.
   d) What is your conclusion? Show both numerical result of the test and a plot demonstrating it.
3. Do students who drink a lot (4 or 5 Dalc) get different final grades in math courses than students who drink less (1–3 Dalc)? (required significance level: 0.05)
   a) Clearly state the null hypothesis and the alternative hypothesis.
   b) Write code to test the hypothesis using simulations. Explain your code.
   c) What is your conclusion? Show both numerical result of the test and a plot demonstrating it.
4. Compute a 95% confidence interval for the median of the first period grade in Portuguese. Explain your code and clearly state the confidence interval (lower and upper values).

   For the following questions, write your answers in text (no code needed).

5. Why is using bootstrap to compute a confidence interval based on a very small sample problematic? Explain in 2-4 sentences.
6. Imagine you are trying to estimate the popularity of different food options on campus. To do this, you stood at the entrance of "Bet Hastudent" and surveyed students about their favorite food options. You collected ratings between 1–5 for each of the options. You then computed a bootstrap confidence interval for the difference between the ratings for "Nuna" (in physics) and Sushi in "Bet Hastudent" with Sushi ratings being higher. Would you trust these results? Explain your answer, use the terminology we learned in class to justify your arguments (3-5 sentences).