Lab 5

Logicals and simple data frames manipulation

Monday Feb. 15, 2016

Due before lab: Mon. Feb. 22, 2016

In this assignment, we work with logical expressions and simple data frames. We work on extracting subsets of the data and putting the commands used inside a function, itself part of a script. Include these scripts (one per task) with your report. There should be one script for each function in this assignment. These scripts should have names such as "task01_yourname.r", etc.; they should be collected into a single archived via tar or zip or gzip and emailed to the TA by the due date along with the report. Answers should be neatly output to the screen by the script.

Task 1: 25 pts

(1) Generate 10000 random even integers tenin the range 2 to 1024. Extract a vector with the random numbers less than 500 and calculate the mean of this vector.

(2) Using the same vector of 10000 random even integers generated in part 1, extract a vector of all integers that are either less than 380 or greater than 882. What is the mean value and standard deviation of the resulting extracted vector? What is the variance?

Task 2: 25 pts

Using the data() function, read the dataset named USJudgeRatings and provide the following information: (In each case, write the command that allowed you to provide the answer to the question.)

(a) What is the class of the datataset

(b) What is the structure of the dataset? How many columns? How many rows in each column?

(c) What is the type of each column? What are the column headers?

(d) Compute the mean and standard deviation of judge integrity for judges with diligence above the mean. Repeat for judges with diligence below the mean. Conclude whether, in your opinion, diligence affects integrity or not. Explain your answer (a simple yes or no is not sufficient .)

(e) Compute the mean and standard deviation of "Familiarity with the law" for judges whose "sound oral rulings" are above the mean. Do the same for judges whose "sound oral rulings" are below the mean. Can you draw any conclusions from this data that seem reasonable?

Task 25 pts

Use the data from "USAccDeaths". What type of data is it? What is its structure? What does this data represent?

Task 4: 25 pts

Read the data.set named esoph.

(0) What is this data about? (Hint. Factors represent categories. You can use factors as if they were strings.

(1) What is the structure of this dataset? What are the types of each column?

(2) Compute the mean number of controls for each age group: "25-34", "35-44", and "45-54".

(3) Compute the mean number of controls for the combination of age groups: "65-74" and "75+".

(4) Compute the total number of cases for all patients age 55 and above.

(5) Create a new dataframe composed of only the columns for Age Group and Tobacco consumption.

For this assignment, the following functions may be useful: (seq(), sample(), mean(), sd(), ?, data.frame(), source(), str(), class(), cat())