Статистика и емпирични методи

Домашно №1

Предаване на хартиен носител до 21 април 2016 г.

В общия случай, ако ни се наложи да изследваме някакви данни, те най-вероятно няма да бъдат включени в пакет на `R`. За да можем да ги обработим тогава, ще трябва да ги прочетем по някакъв начин. Доста често данните се записват в `.csv` (comma separated values) файл. Такива файлове могат да бъдат прочетени в `R` чрез функцията `read.csv()`, на която ѝ е подаден като аргумент стринг с пътя до файла и ѝ е указано дали файлът има header, т.е. дали първият му ред съдържа имената на променливите. Например, ако искаме да прочетем файла `data.csv`, намиращ се на десктопа, в `R` и да го запишем под формата на data frame, трябва да напишем следното в конзолата:

`my_data = read.csv("C:/Desktop/data.csv" , header = TRUE)`.

И така, задачата за домашно е да се изследват данни за 891 пътника на борда на Титаник (източник: kaggle.com). Данните са във файла `titanic.csv`, който е с header, и трябва да бъдат свалени от тук. Записаните променливи са следните:

- `Survived`: Променлива, указваща дали съответният пътник е преживял потъването на Титаник. 0 отговаря на „не", 1 отговаря на „да". Леонардо ди Каприо има 0;
- `Pclass`: Класата, в която е пътувал съответният пътник;
- `Sex`: Полът на пътника;
- `Age`: Възрастта на пътника;
- `SibSp`: Брой братя/сестри и съпрузи/съпруги на съответния пътник на борда;
- `Parch`: Брой родители/деца на съответния пътник на борда;
- `Fare`: Цената на билета на съответния пътник;
- `Embarked`: Мястото, от което пътника се е качил на кораба: C = Cherbourg, Q = Queenstown, S = Southampton.

Задачите към домашното са следните:

- Прочетете данните и ги запишете в data frame в `R`;
- Генерирайте си подизвадка от 500 наблюдения. За целта нека `f_nr` е вашият факултетен номер. Задайте състояние на генератора на случайни числа в `R` чрез `set.seed(f_nr)`. С помощта на *подходяща функция* генерирайте извадка *без връщане* на числата от 1 до 891 като

не забравяте да я запишете във вектор. Използвайте вектора, за да зашишете само редовете със съответните индекси в нов дейтафрейм и работете с него оттук нататък;

- Изчистете данните: за нашите цели ще ни трябват само наблюдения, при които имаме информация за всяка от променливите, но не всеки пътник е споделил каква е възрастта му. Проверете в `R` какво прави функцията `is.na` и я използвайте върху променливата `Age`, за да извикате само редовете, където имаме наблюдения със записана възраст. Запишете резултата в нов дейтафрейм и работете с него оттук нататък;
- Изкарайте на екрана първите няколко (5-6) наблюдения;
- Какъв вид данни (качествени/количествени, непрекъснати/дискретни) са записани във всяка от променливите?
- Изведете дескриптивни статистики за всяка една от променливите;
- Изведете редовете на най-младия и най-стария пътник;
- Изведете редовете на пътниците с цена на билета над 150;
- Какъв е броят на мъжете, отпътували от Queenstown или Southampton, на възраст между 20 и 40 години?
- Направете хистограма на възрастта *само* на пътниците с поне един придружител (брат/сестра/съпруг/съпруга/дете/родител) и нанесете графика на плътността върху нея. Симетрично ли са разположени данните?
- Изследвайте съвместно променливите `Fare` и `Embarked` с подходящ графичен метод. Какво забелязвате? Сравнете извадковите средни и медианите в двете групи и направете извод;
- Изследвайте съвместно променливите `Survived` и `Sex` с честотна таблица и подходящ графичен метод. Коментирайте резултатите;
- Изследвайте съвместно променливите `Age` и `Fare` с подходящ графичен метод. Бихте ли казали, че съществува линейна връзка между тях? Намерете корелацията между величините и коментирайте стойността ѝ.

Инструкции за предаване на домашната работа:

- На първата страница трябва да са написани името, факултетният номер, специалността и административната група;
- Прилага се кодът на `R` и необходимите резултати (вкл. графики), както коментари и интерпретация на получените статистически резултати;
- Максимален обем: 5 листа;
- Бонус точки при двустранно принтиране ;)