Статистика и емпирични методи

Домашно №1

Предаване до 2016-11-21 10:00:00

Данните за домашното са записани 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 705 вида покемони. Данните са във файла `pokemon.csv`, който е с header, и трябва да бъдат свалени от тук. Записаните променливи са следните:

- `Number`: Пореден номер на вида покемон;
- `Name`: Името на вида покемон;
- `Type1`: Първичен тип на вида покемон (всеки вид има първичен тип);
- `Type2`: Вторичен тип на вида покемон (не всеки вид има вторичен тип);
- `Attack`: Точки, определящи силата на атаката;
- `Defense`: Точки, определящи защитната способност;
- `Height`: Средна височина на вида покемон;
- `Weight:` Средно тегло на вида покемон.

Задачите към домашното са следните:

- Прочетете данните и ги запишете в data frame в `R`;
- Генерирайте си подизвадка от 600 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 705 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;
- 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-високия и на най-лекия покемон;
- Изведете редовете на покемоните с общ брой точки за атака и защита над 220;
- Колко на брой покемони имат първичен или вторичен тип "Dragon"или "Flying"и са високи над един метър?
- Направете хистограма на теглото *само* на покемоните с вторичен тип и нанесете графика на плътността върху нея. Симетрично ли са разположени данните?
- За покемоните с първичен тип "Normal"или "Fighting"изследвайте съвместно променливите `Type1` и `Height` с подходящ графичен метод. Забелязвате ли outlier-и? Сравнете извадковите средни и медианите в двете групи и направете извод;
- Изследвайте съвместно променливите `Height` и `Weight` с подходящ графичен метод. Бихте ли казали, че съществува линейна връзка между тях? Намерете корелацията между величините и коментирайте стойността ѝ. Начертайте регресионна права (линейната функция, която най-добре приближава функционалната зависимост). Ако е наблюдаван нов вид покемон с височина 2.1 метра, какво се очаква да е теглото му на базата на линейния модел?

Инструкции за предаване на домашната работа:

- Предаването на домашното ще бъде през страницата на курса в moodle;
- Домашното трябва да е в .pdf формат. 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.