Статистика
R

Защо R?

- безплатен
- open-source
- лесен синтаксис
- удобен

Характеристики на R

Обектно-ориентиран

Динамичен

Поддържа type inference

Някои основни типове

- character - - “statistics” (да, на цялото нещо му казват character)
- numeric
  - за да укажем изрично, че е integer, се добавя L накрая. 1L
- logical - TRUE, FALSE
- complex - 1 + 4i
- NA - при липса на данни
- NaN

typeof()

Напишете няколко примера и проверете типа им.

Структури от данни

- вектор
- списък
- матрица
- data frame
- factor
- table

Ние ще използваме главно vector и data frame.

Вектори

- ще ги използваме най-много
- ще ги конструираме с функцията **c**
  - c – комбинира данни във вектор или лист.
- индексиране
  - кои индекси да вземе
- Може ли в R векторът да има елементи от различни типове? Т.е. да са нехомогенни? Не!

```
x = c(1, 2, 3)

length(x)

1:10

# Което е синтактична захар за:

seq(1, 10, by = 1)

x[1:6]
```

x[c(индексите, които искаме от вектора)]

С минус означаваме кои индекси да пропусне:

x[-c(1, 6)]

x[c(вектор от TRUE и FALSE кои елементи да се вземат)]

which(x == 1) – връща индексите на елементите в x, които връщат TRUE за условието. После можем да вземем самите елементи от x така: x[which(x == 1)].

От друга страна можем да получим масив от TRUE и FALSE x == 1 и за да вземем елементите: x [x == 1].

Обединение: x[x< -2 | x > 2]

Сечение на условия: x[x> -2 & x < 2]

Операторът + при булевите стойности е реализиран така:

TRUE + TRUE == 2

Затова:

sum(x==1) # броят на елементите в x, които са равни на 1.

Защо?

Покомпонентно се извършват операциите: + * -, когато са приложени над вектори.

Ако два вектора не са с еднаква дължина, R повтаря по-краткия:

c(1, 2, 3, 4, 5) + c(10, 100)

Не пази референции, а самите стойности!

Задачи

https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

11стр., задача 2.2, 2.5

Полезно: scan()

Какво са?

mean (средно), median, mode, min, max, var (дисперсия), range (размах), sd (стандартно отклонение)

$$\mathrm{SD}(X) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Функциите им в R

- mean
- median
- min
- max
- range (връща вектор)
- var
- sd

Задачи

https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

Задача 2. 6

Матрици

- няма да се ползват в курса
- да знаете, че съществуват
- създаване
- индексиране
  - ред
  - колона
  - елемент
- dimnames
  - list от вектори първо с имената на редовете, после на колоните

```
m = matrix(nrow = 2, ncol = 2)
m = matrix(1:6, nrow = 2, ncol = 3)
# запълват се вертикално, а ако искаме хоризонтално, параметъра  # byrow = TRUE, който по подразбиране е FALSE.
mdat = matrix(c(1,2,3, 11,12,13), nrow = 2, ncol = 3, byrow = TRUE, dimnames = list(c("row1", "row2"), c("C.1", "C.2", "C.3")))
```

Data Frame

- извадки от данни
- можем да ги прочетем примерно от csv
  - read.csv()
- Общо взето е множество от наименувани вектори с еднаква дължина.
- df = data.frame(c(1,2), c(3,4))

Полезни функции с Data Frames

- head() - дава първите 6 реда
- tail() - дава последните 6 реда
- dim() - размерности (вектор от брой редове и брой колони)
- nrow()
- ncol()
- names() -дава имената на колоните (работи не само с df)