ლაბორატორია 2: შესავალი R-გარემოში

დავით სიჭინავა

23 მარტი 2018 წ.

საკითხები

- გარე ფაილების წაკითხვა
- მარტივი სიხშირის ცხრილები
- არითმეტიკული ოპერაციები

ინსტრუქციები:

თანმიმდევრობით შეასრულეთ მითითებული ამოცანები. თქვენს .rmd ფაილს სახელწოდება მიანიჭეთ შემდეგი ფორმით: თქვენი გვარი_lab2.rmd. მაგალითად:

```
sichinava_lab2.Rmd
```

დავალება:

გარე ფაილების წაკითხვა R-ში

პირველ რიგში, შექმენით სამუშაო ფოლდერი მეორე ლაბორატორიისთვის და დაარქვით სახელი, მაგალითად - lab_2.

გახსენით 'R-studio' და შექმენით ახალი R-ბლოკნოტი. დაარქვით სახელი და შეინახეთ ფოლდერში lab_2.

$R-$-ის უპირატესობა ისაა, რომ მასში შესაძლებელია მონაცემთა ბევრი ფორმატის გახსნა, მაგალითად MS Excel-ის, ტექსტური ფაილების, Stata, SPSS, SAS, SQL ცხრილების, სივრცით მონაცემთა (śhp), „ჯავასკრიპტის"ნოტაციის დოკუმენტების (json) და ა.შ. ფაილები შესაძლოა, შენახული იყოს როგორც თქვენს კომპიუტერში, ასევე - „სადმე" ქსელში, ინტერნეტის ჩათვლით.

როგორც წესი, მონაცემები ცხრილურ (ტაბულარულ) ფორმატში ინახება, სადაც თითოეული რიგი წარმოადგენს *შემთხვევას*, ხოლო სვეტი - *ცვლადს*. ასევე არსებობს მონაცემთა შენახვის და ორგანიზების მრავალი ფილოსოფია (e.g. tidy მონაცემები,

დოკუმენტური ბაზები, რელაციური ცხრილები), თუმცა მათ შესახებ დეტალურად ამ კურსის ფარგლებში არ ვისაუბრებთ.

R-ის საბაზო ფუნქციები კარგად უმკლავდება ტაბით და მძიმით გამოყოფილ მონაცემებს. როგორც წესი, ჩვენი მონაცემები შენახულია ტექსტურ ფაილებს, რომლებსაც *.csv*, *.txt* ან *dat* გაფართოება გააჩნიათ. პრაქტიკაში ისინი შემდეგნაირად გამოიყურებიან:

```
###

1   6   a
2   7   b
3   8   c
4   9   d
5   10  e

###

v1,v2,v3
1,2,3
4,5,6
7,8,9
a,b,c
```

მუშაობის დასაწყებად თქვენს ბლოკნოტში ჩასვით კოდის „ნაგლეჯი" (Code chunk). „ნაგლეჯში" აქცენტის ნიშნებს შორის ჩაწერეთ კოდი, რომელიც სამუშაო დირექტორიად მიუთითებს _lab2.rmd-ს. ამ ლინკიდან გადმოიწერეთ $Galton.csv$ ფაილი და შეინახეთ თქვენს სამუშაო დირექტორიაში. $Galton.csv$ *მძიმით გამოყოფილი* ფაილია, რაშიც ადვილად დარწმუნდებით, თუ ფაილს Notepad-ით გახსნით. csv ფაილების წასაკითხად უნდა გამოიყენოთ $read.csv$ ფუნქცია, კერძოდ:

```
galton <- read.csv("Galton.csv")
```

ამ სინტაქსში პირველი $galton$-ი წარმოადგენს R-ის ახალი *ობიექტის* სახელწოდებას, რომელსაც მოგვიანებით გამოთვლებისას მივმართავთ. ფუნქცია არის $read.csv$. ყურადღება მიაქციეთ, რომ ფაილის სახელები მითითებული უნდა იყოს ან ბრჭყალებში, ან - აპოსტროფებში.

სერ ფრენსის გალტონი, ცნობილი ბრიტანელი სტატისტიკოსი და შეთავსებით, რასისტი და ევგენიკოსი, დაინტერესებული იყო, თუ რა გავლენას ახდენს წინაპრების ფიზიოლოგიური მახასიათებლები შთამომავლობაზე. გალტონმა შეაგროვა დაახლოებით ორასამდე ოჯახის მონაცემები და შეისწავლა, თუ რა გავლენას ახდენდა შვილების სიმაღლეზე მშობლების ფიზიკური განზომილებები.

სანამ მონაცემთა ანალიზს დავიწყებთ, ჯერ ვნახოთ, თუ რას წარმოადგენს ჩვენი მონაცემები. გამოიყენეთ ფუნქცია $names$ იმის სანახავად, თუ რა სახელწოდების სვეტებია წარმოდგენილი ცხრილში. რამდენი ცვლადია $galton$ ცხრილში? ჩაინიშნეთ მათი სახელწოდებები თქვენს ბლოკნოტში.

```
names(galton)
```

მარტივი აღწერითი სტატისტიკა R-ში

მშვენიერია. მოდი, განვიხილოთ ცვლადი $height$. იმისთვის, რათა R-ში ობიექტს მივმართოთ, საჭიროა დოლარის ნისნის გამოყენება. ქვემოთ მოყვანილი სინტაქსი ცვლადს კონსოლში გამოიტანს:

```
galton$height
```

მოდი, გამოვთვალოთ ცვლად $height$-ის საშუალო, მედიანა და სტანდარტული გადახრა:

```
mean(galton$height)

median(galton$height)

sd(galton$height)
```

დაწერეთ კოდი, რომელიც გამოთვლის საშუალოს, მედიანას და სტანდარტულ გადახრას ცვლადებისთვის $father$ და $mother$. Write a very simple description of the dataset.

მონაცემთა ტრანსფორმაცია R-ში

როგორც ხედავთ, ბაზაში სიმაღლეები მოცემულია დუიმებში. ერთ დუიმში 2.54 სანტიმეტრია. შექმენით ახალი ცვლადი $height\_cm$ სადაც გაზომილი იქნება შვილების სიმაღლე სანტიმეტრებში:

```
galton$height_cm <- galton$height*2.54
```

შეაჯამეთ ახალი ცვლადი და გამოთვალეთ მისი საშუალო, მედიანა და სტანდარტული გადახრა.

ანალოგიურად შექმენით ახალი ცვლადები, რომლებიც სანტიმეტრებში გაზომავენ მამის და დედის სიმაღლეებს. გამოთვალეთ ამ ახალი ცვლადების საშუალო, მედიანა და სტანდარტული გადახრა.

ხშირ შემთხვევაში, გვჭირდება, დავაჯგუფოთ ცვლადის მნიშვნელობები, მაგალითად - ასაკის ნაცვლად, შევქმნათ ახალი ცვლადი კატეგორიებით (მაგ. მილენიალები, ბეიბი-ბუმერები და ა.შ.). მოდი, დავაგენერიროთ შვილის სიმაღლის აღმნიშვნელი ახალი ცვლადი, რომელსაც ექნება ოთხი მნიშვნელობა: 1, რომელიც შეესაბამება 150 სანტიმეტრზე დაბალ ბავშვებს, 2 - 150-169 სანტიმეტრი სიმაღლის, 3 - 170-189, ხოლო 4 - ყველაზე მაღალ (190+) ჯგუფს. ახალ ცვლადს დავარქვათ $height\_gr$.

```
galton$height_gr <- galton$height_cm
```

ძალიან კარგი. ახლა R-ს უნდა ვუთხრათ, რომ $height\_gr$ ჩაანაცვლოს 1-ით, სადაც $height\_cm$ ნაკლებია 150-ზე:

```
galton$height_gr[galton$height_cm < 150] <- 1
```

როგორც ხედავთ, კვადრატული ფრჩხილები მიუთითებენ იმ კოდზე, რომელიც ფილტრავს მონაცემებს. აქ შესაძლებელია, გვქონდეს რთული სინტაქსიც:

```
galton$height_gr[galton$height_cm>=150 & galton$height_cm<170 ] <- 2

galton$height_gr[galton$height_cm>=170 & galton$height_cm<190 ] <- 3

galton$height_gr[galton$height_cm>=190] <- 4
```

ადვილია? მონაცემებს ასევე შეიძლება, მივანიჭოთ კატეგორიის ტექსტური განმარტებები. ამისთვის, რაოდენობრივი ცვლადი ფაქტორულად უნდა ვაქციოთ:

```
galton$height_gr <- factor(galton$height_gr, labels=c("Short", "Medium", "Tall", "Very tall"))
```

მოდი, იმავე მიდგომით დააჯგუფეთ დედის სიმაღლეც. ამასთან, გაითვალისწინეთ, რომ ეს ცვლადიც სანტიმეტრებშია გაზომილი.

Frequency tables in R

გამოთვალეთ, რამდენი დაბალი, საშუალო, მაღალი და ძალიან მაღალი ადამიანია გალტონის ბაზაში. ამისთვის, უნდა შევქმნათ ცვლად $height\_gr$ სიხშირის ცხრილი:

```
table(galton$height_gr)
```

ანალოგიურად, შექმენით დედის სიმაღლის კატეგორიების სიხშირის ცხრილი.

ავაჰმე. მზადაა. როგორ ჩავაბარო?!

დაზიპეთ ფოლდერი და დაარქვით სახელი შემდეგი ფორმატით: $surname\_lab2.zip$.

```
shubladze_lab2.zip
```

ატვირთეთ თქვენი დავალება ამ ლინკზე მომდევნო შეხვედრის დაწყებამდე. წარმატებები!