Домашна работа за СЕМ-Практикум

Божидара Петрова Пачилова, ф.н. 62172

спец. Софтуерно инженерство, курс 3, група 1

# **Анализ на данни**

За целите на домашната работа, направих [анкета (линк към нея)](https://www.surveymonkey.com/r/969XSVG) на тема „Седмична спортна активност“. Анкетата събира следните данни:

* Пол
* Възраст
* Брой тренировъчни дни на седмица
* Вид спорт

Анкетата попълниха 70 души. Резултатите може да се видят в прикачената таблица sports.xlsx.

## Изследване на едномерните променливи

> library(readxl)

> sports <- read\_excel("sports.xlsx")

1. Разпределение на участниците по пол в проценти:

F M

60 40

1. По възраст:

> hist(sports$age, xlab = "age", ylab = "percentage", main = "Histogram of age", col=rainbow(6), prob=T)

> rug(jitter(sports$age))



Min. 1st Qu. Median Mean 3rd Qu. Max.

16.00 21.00 21.00 25.41 23.00 72.00

> sd(sports$age)

[1] 10.8287

1. Брой на нетрениращите:

> length(which(sports$training\_frequency == 0))

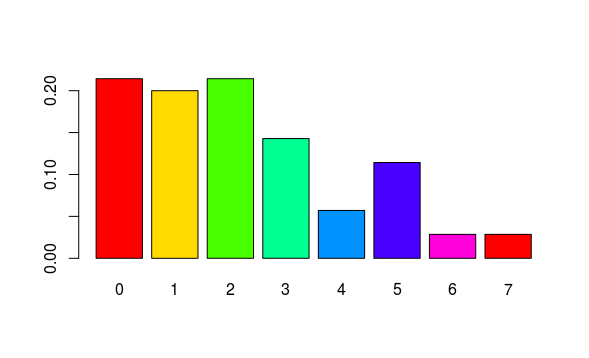
[1] 15

т.е. 15 от 70 души не спортуват или 21% от всички.

1. Визуализация на честота на спортуването в проценти от хората

> tt2 <- table(sports$training\_frequency)

> barplot(prop.table(tt2), col = rainbow(7))



Както виждаме, най-много преобладават неспортуващите (тези, които спортуват 0 дни) и спортуващите по 1-2 пъти седмично.

Разпределението не може конкретно да се определи – има три пика.

> summary(sports$training\_frequency)

Min. 1st Qu. Median Mean 3rd Qu. Max.

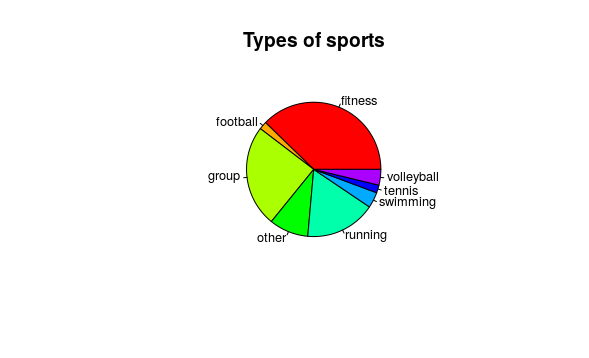
0.000 1.000 2.000 2.229 3.000 7.000

1. Разпределение на видовете спорт

> sports.table <- table(sports$sport)

> pie(sports.table[-4], main = "Types of sports", cex = 0.8, col = rainbow(length(sports.table)))

\*sports.table[-4] се използва, за да се премахне категория “none”, въведена от нетрениращите.



Виждаме, че фитнесът е най-разпространената форма на физическа активност, следван от груповите занимания (Танци, Йога, Пилатес и всички др.). Тичането също е популярно, докато спортове като футбол, волейбол, плуване и тенис в тази извадка бяха посочени от 1-2 човека.

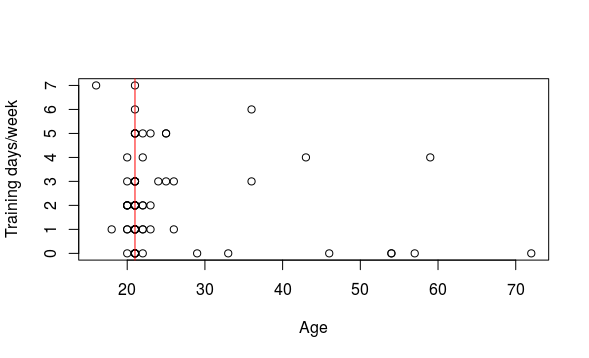
Категория „други“ включва най-различни спортове, като виждаме, че не малка част са посочили такива. Сред тях бяха: фехтовка, канадска борба, фризби и електронни спортове.

## Изследване на двумерни данни

1. **Честота на тренировките според възрастта**

> plot(sports$age, sports$training\_frequency, xlab = "Age", ylab = "Training days/week")

> abline(v = median(sports$age), col="red")

Виждаме, че резултатите не са никак детерминирани – нито може да се каже, че 20-годишните тренират, а възрастните – не, нито обратното.

Сега формално изследваме корелацията между двете променливи:

> round(cor(sports$age, sports$training\_frequency, method = "pearson"), 2)

[1] -0.18

Коефициентът, изчислен по метода на Pearson, попада в скалата за слаба корелация.

Използвайки методите на Spearman и Kendall, корелацията също излиза много слаба:

> round(cor(sports$age, sports$training\_frequency, method = "spearman"), 2)

[1] -0.06

> round(cor(sports$age, sports$training\_frequency, method = "kendall"), 2)

[1] -0.04

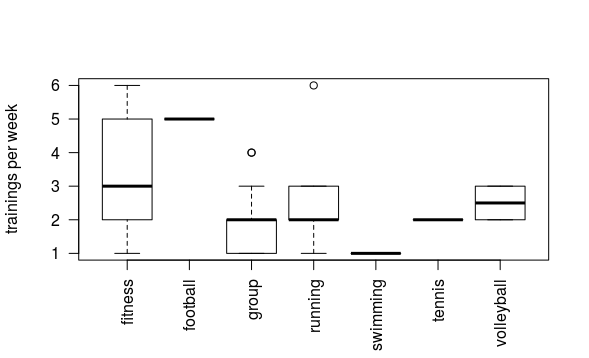
**Извод:** Противно на очакваното, изглежда, че липсва стойностна връзка между възрастта и честотата на трениране. Предполагам, че това се дължи и на факта, че извадката не е достатъчно представителна.

1. **Честота на трениране според спорта**

Първо махаме неспортуващите и избираме само категориите, различни от „други“ и неспортуващите:

> valid\_categories <- sports[which(sports$sport != "other" & sports$sport != "none"),]

> boxplot(valid\_categories$training\_frequency ~ valid\_categories$sport, las=2, ylab="trainings per week", xlab="")



**Изводи:**

* Фитнес се тренира в целия интервал от възможен брой дни, но най-много по 3 пъти седмично;
* Групови занимания се тренират 1-2 пъти на седмица – очакван за мен резултат (и има един outlier с 4 пъти);
* Тичането е 2-3 пъти на седмица (и има един екстремист, тичащ почти всеки ден);

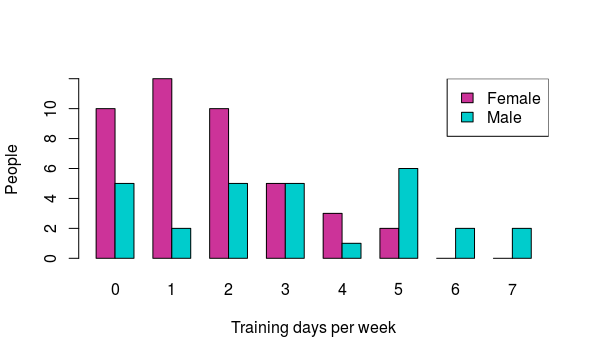
1. **Честота на трениране по пол**

> barplot(table(sports$gender, sports$training\_frequency), col=c("#CC3399","#00CCCC"), beside = T, xlab = "Training days per week", ylab = "People")

> legend("topright",

+ legend = c("Female", "Male"),

+ fill = c("#CC3399","#00CCCC"))



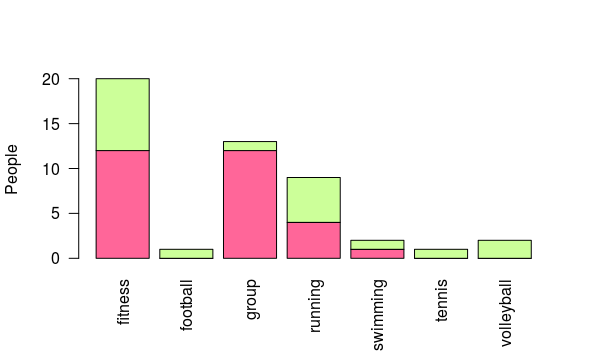
**Извод:** От графиката разбираме няколко неща:

* Жените, които не тренират са повече от мъжете, които не тренират (0 дни/седмица);
* Ако тренират, жените го правят най-често веднъж или два пъти в седмицата;
* За мъжете не може да се забележи зависимост в броя на тренировъчните дни, но се вижда, че при тях се достигат екстремните стойности – 6, 7 тренировки/седмица.
* Като цяло мъжете повече се посвещават на тренировки – 5 дни в седмицата е най-разпространената честота.

1. **Видове спорт по пол**

> barplot(table(valid\_categories$gender, valid\_categories$sport), col= c("#FF6699","#CCFF99"), beside = F, las = 2, ylab = "People")

> legend("topright", c("Female", "Male"), fill = c("#FF6699","#CCFF99"))



**Извод:**

* Излиза, че жените ходят на фитнес повече от мъжете, но това най-вероятно се дължи и на факта, че повече жени са попълнили анкетата.
* Жените тренират най-много групови занимания (заедно с фитнеса).
* Мъжете тренират разнообразни спортове.

1. Изследването на вид спорт по възраст не би имало смисъл в този анализ.

**Заключение:**

Целите на анализа на получените данни бяха:

* Да разбера колко от познатите ми спортуват и какво;
* Да проверя дали има зависимост между различните критерии.

Резултатите са, че:

* Повечето хора спортуват.
* Няма значима връзка между честота на трениране и възраст (поне при тези данни).
* Мъжете като цяло тренират по-често и повече от жените.
* Мъжете тренират по-широка гама от спортове.
* Фитнесът е спорта, който показва най-много постоянство на седмично ниво.