Курсов проект по Статистика

Тема: Зависимост между приема на кофеин и съня на хората

Изготвил:

Иван Тосков

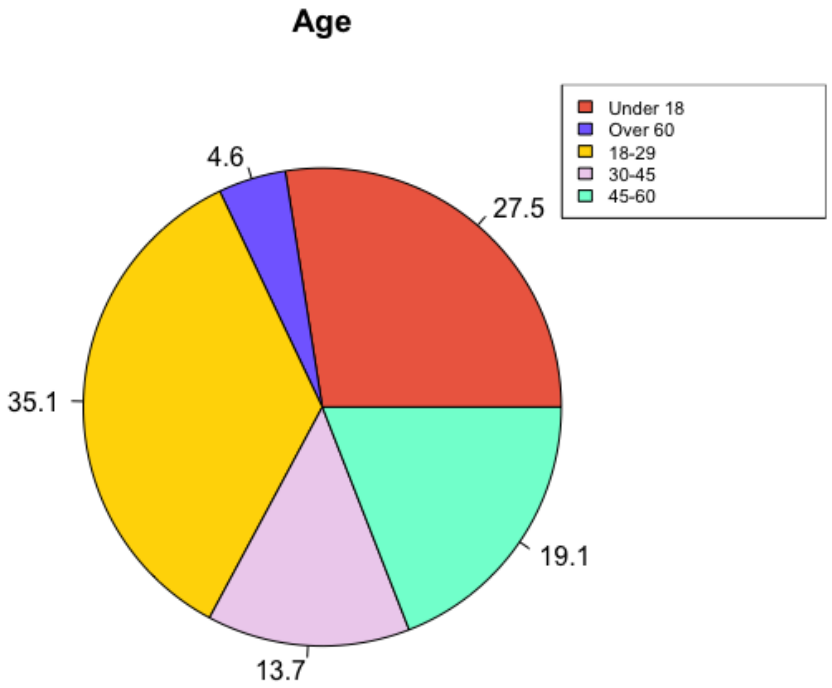
ИС, 2 курс, група 3

ФН: 71953

Направих анкета относно приема на кофеин при хората, като целта и е да разбера как кофеина влияе върху съня им.
Анкетата се състои от 7 въпроса, като участие в нея взеха 131 човека.
В анкетата се съдържат 5 категорийни и 2 числови променливи.

1.Анализ на едномерни променливи

Въпрос 1: Каква е вашата възраст? – категорийна величина



От кръговата диаграма се вижда, че най-голяма част от анкетираните са на възраст между 18 и 29 години, следвани от хора на възраст под 18 години. Най-малка част от анкетираните са хора над 60 години.

Въпрос 2: Консумирате ли кофеинови напитки? – категорийна величина

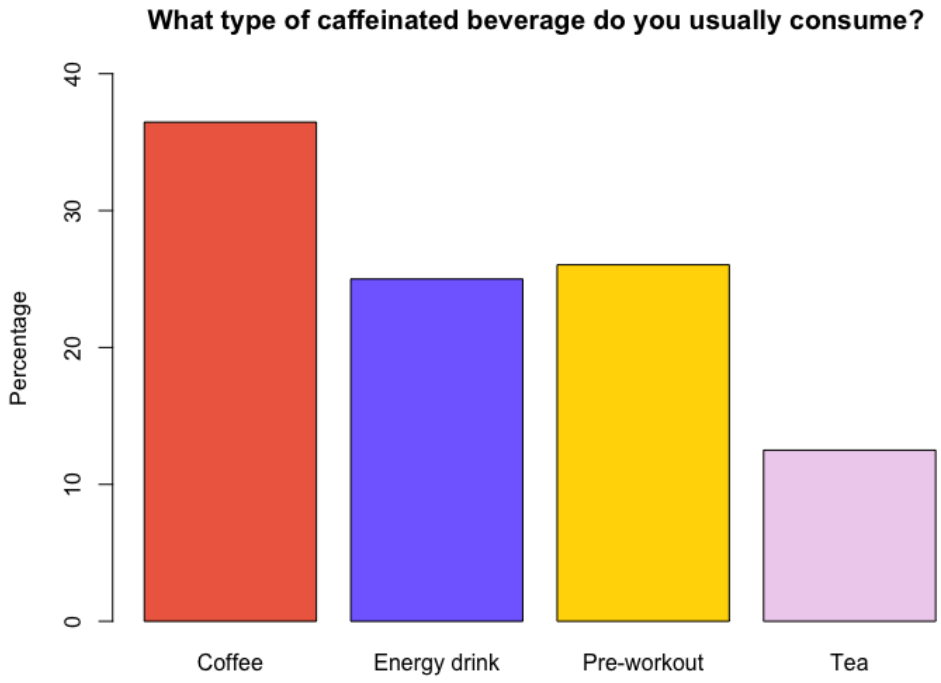
```
> # Q3: Do you consume caffeinated drinks?
> consumationTable <- round(prop.table(table(caffeineConsumation)) * 100)
> consumationTable
caffeineConsumation
 No Yes
 27  73
```

Отговорите на този въпрос намираме като използваме процентна таблица в R.
На нея се вижда, че 73% от анкетираните консумират кофеинови напитки, а 27 – не.

Въпрос 3: Каква кофеинови напитки консумирате обикновено? – категорийна величина

```
# Q4: What type of caffeinated beverage do you usually consume? - Categorical data
caffeineBeverages[caffeineBeverages != ""]
beveragesTable <- table(caffeineBeverages[caffeineBeverages != ""])
beverages <- c("Coffee", "Energy drink", "Pre-workout", "Tea")
barplot(round(prop.table(beveragesTable)*100, 2), col = myColors,
        main = "What type of caffeinated beverage do you usually consume? ",
        ylim = c(0, 40), ylab = "Percentage", names.arg = beverages)
```

Използвам barplot, за да представя графично честотното разпределение.



На графиката ясно се вижда, че най-голяма част от анкетираните(приемащи кофеинови напитки) предпочитат кафе, а най-малко – чай.

Въпрос 4: Приблизително колко кофеинови напитки приемате дневно? – числова променлива

Намирам честотното разпределение на числовата променлива като използвам таблица.

```
> dailyCaffeineTable <- table(dailyCaffeineBeverages)
> dailyCaffeineTable
dailyCaffeineBeverages
 0  1  2  3  4  5
35 46 20 21  6  3
```

Намирам модата(най – често срещаната стойност) и установявам, че най-голяма част от приемащите кофеин анкетирани, консумират по 1 кофеинов продукт дневно.

```
> names(dailyCaffeineTable)[dailyCaffeineTable == max(dailyCaffeineTable)]
[1] "1"
```

Функцията summary ни дава минималната стойност, първия квартил, медианата, средната стойност, третия квартил и максималната стойност.

```
> summary(dailyCaffeineBeverages)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   1.000   1.435   2.000   5.000
```

Вариация

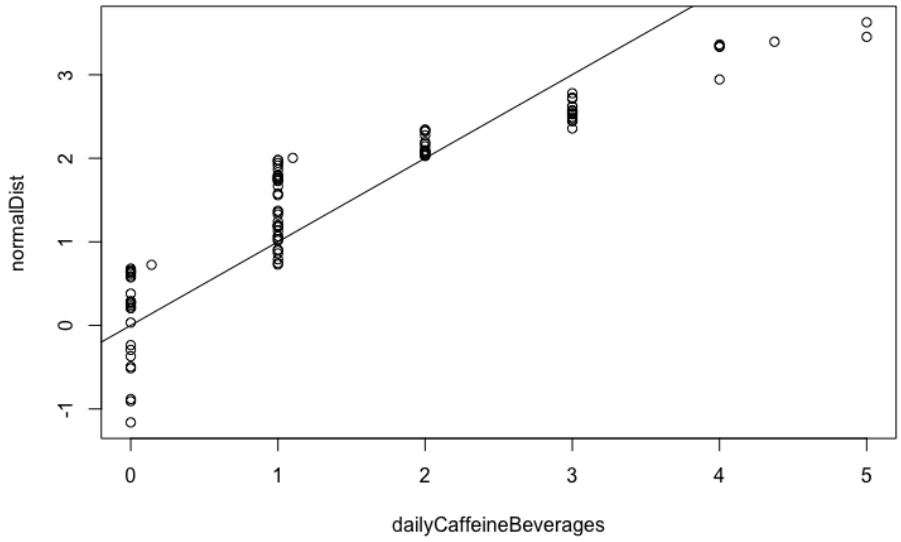
```
> var(dailyCaffeineBeverages, na.rm = TRUE)
[1] 1.663065
```

Стандартно отклонение

```
> sd(dailyCaffeineBeverages, na.rm = TRUE)
[1] 1.289599
```

Проверка за нормално разпределение с помощта на qqplot

```
normalDist <- rnorm(10^2, mean=mean(dailyCaffeineBeverages, na.rm = TRUE),
                    sd=sd(dailyCaffeineBeverages, na.rm = TRUE))
qqplot(dailyCaffeineBeverages, normalDist)
abline(a=0, b=1)
```



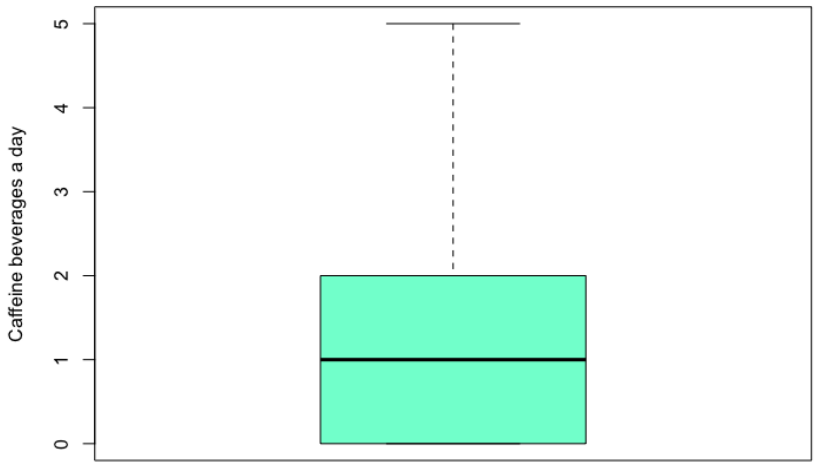
От графиката ясно се вижда, че нямаме нормално разпределение.

```
        Shapiro-Wilk normality test

data:  dailyCaffeineBeverages
W = 0.87241, p-value = 3.084e-09
```

Изпълняваме Shapiro-Wilk normality test и получаваме, че p-value = 3.084e-09 < 0.05, което означава, че нямаме нормално разпределение.

Правя проверка за потенциални outlier-и като използвам boxplot.



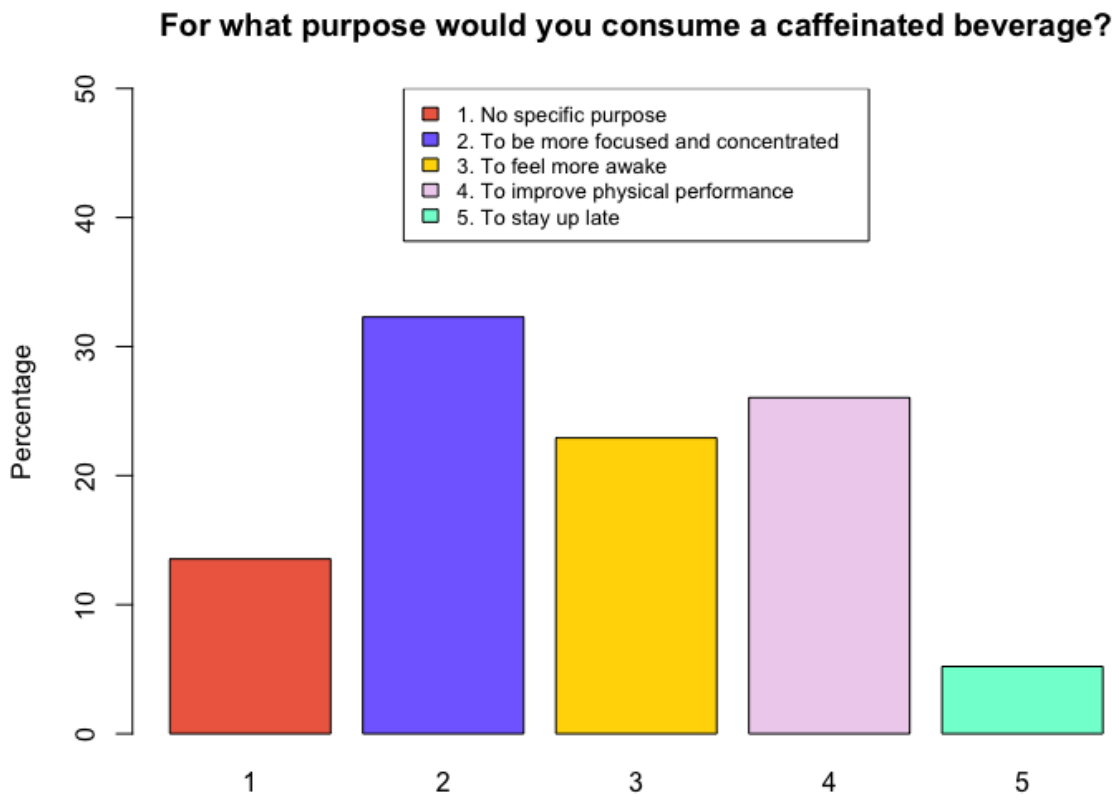
Установявам, че няма потенциални outlier-и.

Въпрос 5: С каква цел консумирате продукти съдържащи кофеин? – категорийна променлива

```
> purposeTable <- table(purpose)
> round(prop.table(purposeTable)*100, 2)
purpose
                No specific purpose To be more focused and concentrated
                              13.54                                32.29
                 To feel more awake     To improve physical performance
                              22.92                                26.04
                    To stay up late
                               5.21
```

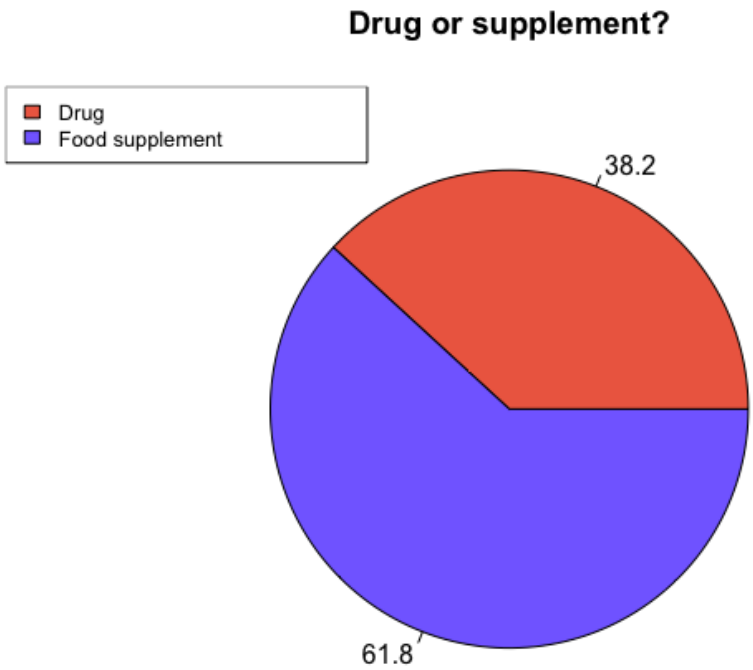
```
row.names(purposeTable) <- paste(1:nrow(purposeTable))
barplot(round(prop.table(purposeTable)*100, 2), col = myColors,
        main = "For what purpose would you consume a caffeinated beverage?",
        ylim = c(0, 50), ylab = "Percentage")
legend(x = "top",
       legend = c("1. No specific purpose", "2. To be more focused and concentrated",
                                  "3. To feel more awake", "4. To improve physical performance",
                                   "5. To stay up late"),
       cex = 0.8, text.width = 2.5, fill = myColors)
```

С помощта на barplot представям честотното разпределение.



От диаграмата се разбира, че най-голяма част от анкетираните(консумиращи кофеин) приемат кофеинови напитки за да се чувстват по фокусирани и концентрирани - 32 % и тези, които искат да подобрят физическото си представяне(спортисти) – 26%.
Най-малка част приемат кофеин за да стоят до по-късно вечер.

Въпрос 6: Според вас, кофеинът наркотик ли е или хранителна добавка? – категорийна променлива



На кръговата диаграма ясно се вижда, че според повечето анкетирани кофеинът е хранителна добавка.

Въпрос 7: По-колко часа спите всяка нощ? – числова променлива

Намирам честотното разпределение на числовата променлива като използвам таблица

```
> hoursOfSleepTable <- table(hoursOfSleep)
> hoursOfSleepTable
hoursOfSleep
 4  5  6  7  8  9 10 11
 1  1 18 44 51 10  4  2
```

Намирам модата(най – често срещаната стойност) и установявам, че най-голяма част от анкетираните, спят по 8 часа на вечер.

```
> names(hoursOfSleepTable)[hoursOfSleepTable == max(hoursOfSleepTable)]
[1] "8"
```

Функцията summary ни дава минималната стойност, първия квартил, медианата, средната стойност, третия квартил и максималната стойност.

```
> summary(hoursOfSleep)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.000   8.000   7.519   8.000  11.000
```

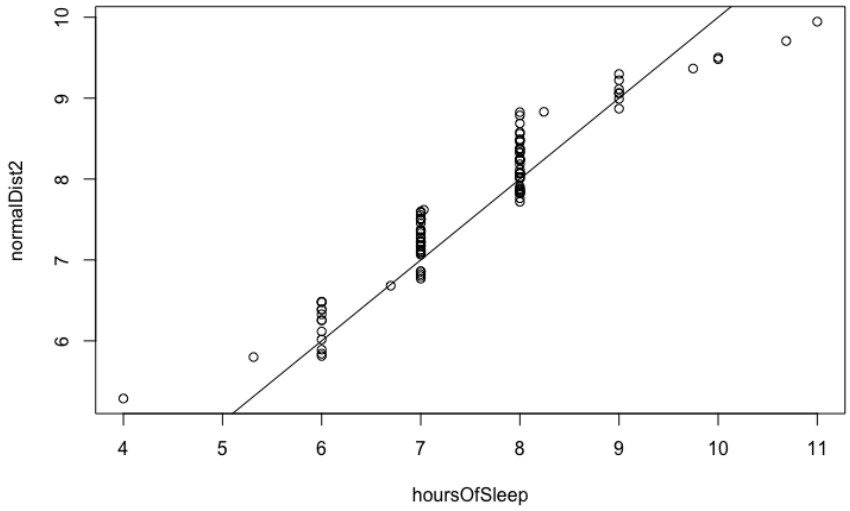
Вариация

```
> var(hoursOfSleep, na.rm = TRUE)
[1] 1.190018
```

Стандартно отклонение

```
> sd(hoursOfSleep, na.rm = TRUE)
[1] 1.090879
```

Проверяваме за нормално разпределение с помощта на ggplot

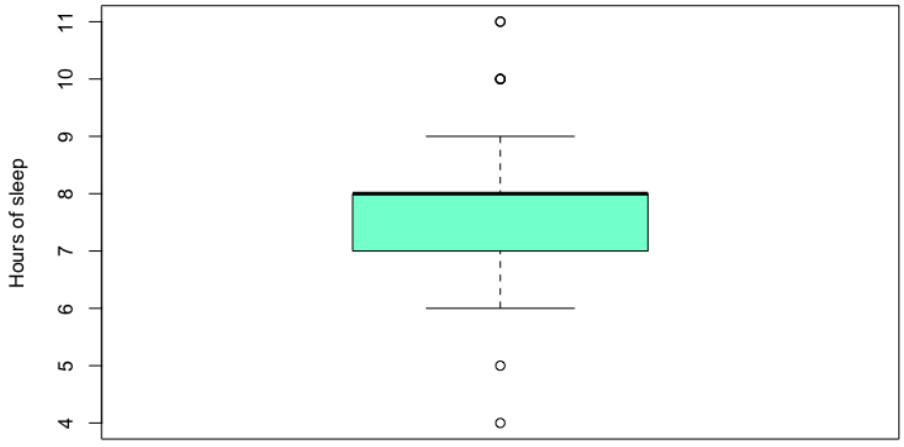



```
        Shapiro-Wilk normality test

data:  hoursOfSleep
W = 0.90325, p-value = 1.055e-07
```

От графиката и от Shapiro-Wilk normality test, където p-value = 3.084e-09 < 0.05, правим заключение, че нямаме нормално разпределение.

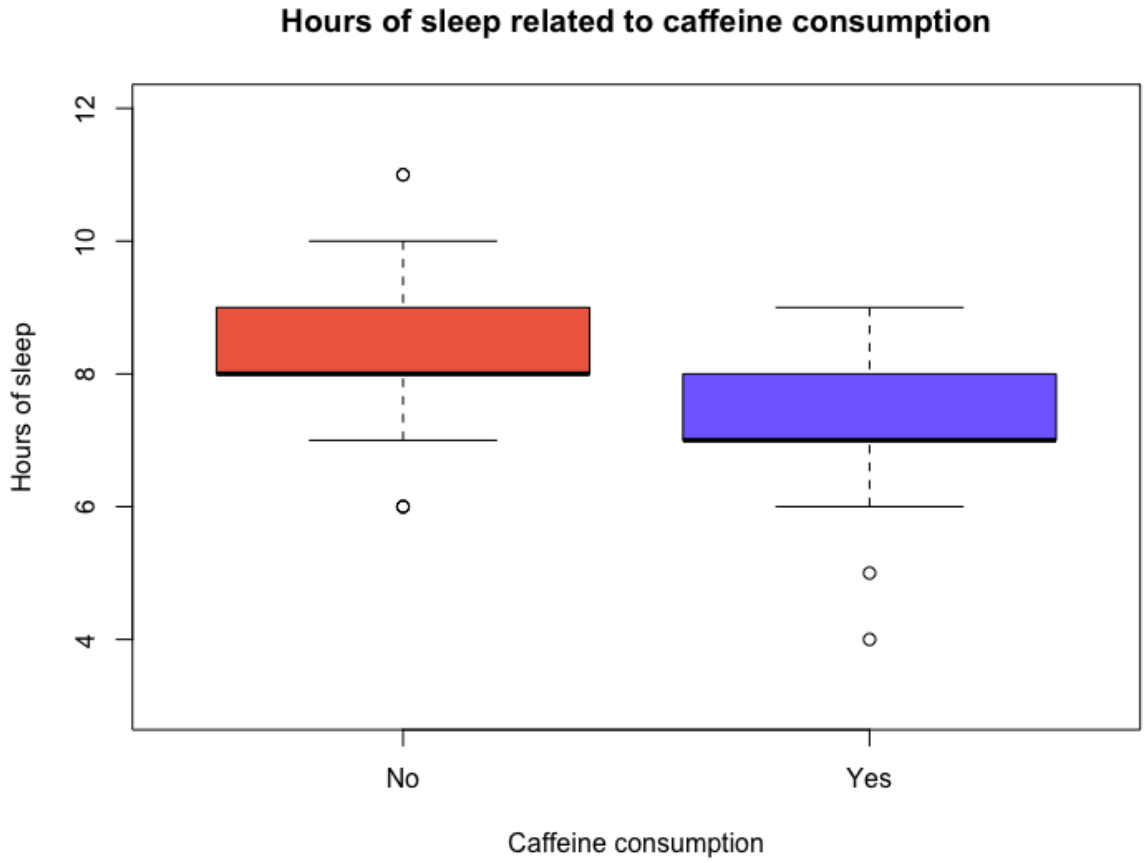
Проверка за потенциални outlier-и с помощта на boxplot. – Имаме потенциални outlier-и



2. Анализ на многомерни променливи

Категорийна vs Числова

Ще разгледам връзката между консумацията на кофеин и часовете сън всяка нощ. За тази цел, ще използвам boxplot.



Удебелената черта представлява медианата. От двете страни на медианата са първия и третия квартил. И в двата случая медианата съвпада с първия квартил. Дължините на опашките са минималната и максималната стойност.
От графиката разбираме, че има разлика при часовете сън между хората консумиращи кофеин и тези които не консумират.

```
        Shapiro-Wilk normality test

data:  consumers
W = 0.88262, p-value = 3.676e-07

> shapiro.test(nonComsumers) # is not normal dist

        Shapiro-Wilk normality test

data:  nonComsumers
W = 0.89663, p-value = 0.003213
```

Правя тест за нормално разпределение на часовете сън при консуматорите и не-консуматорите и установявам, че и в двата случая няма нормално разпределение. Поради тази причина ще ползвам Wilcoxon rank sum test.

Хипотези:
H0: Няма разлика в часовете сън
H1: Има разлика

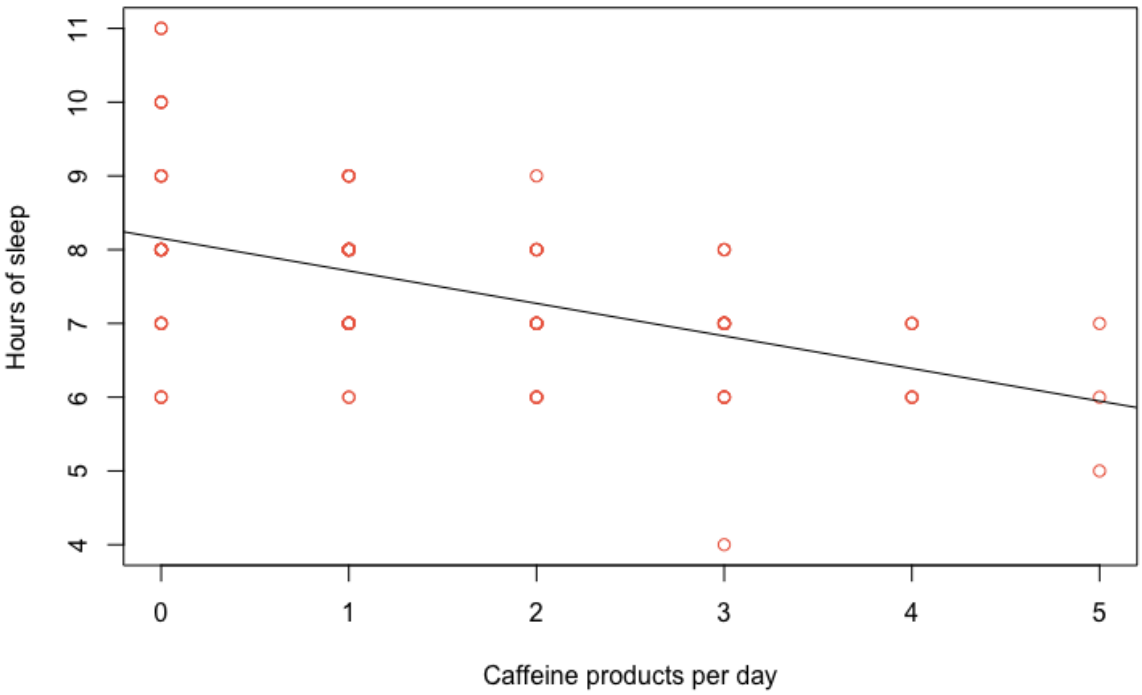
```
        Wilcoxon rank sum test with continuity correction

data:  hoursOfSleep by caffeineConsumation
W = 2400, p-value = 7.987e-05
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 2.570046e-05 1.000079e+00
sample estimates:
difference in location
             0.9999701
```

p-value има стойност по-малка от 0.05, така че отхвърляме H0.
Установявам, че има разлика в часовете сън през нощта между групата която консумира кофеинови продукти и тази която не консумира.

Числова vs Числова

Ще разгледам зависимостта между броя консумирани кофеинови напитки на ден и часовете сън през нощта.



От графиката разбираме, че има отрицателна линейна връзка

```
> rho <- round(cor(dailyCaffeineBeverages, hoursOfSleep), 3)
> rho
[1] -0.523
```

Има обратна корелация между приема на повече кофеинови продукти и часовете сън. Това означава, че колкото повече кофеинови продукти консумират анкетираните, толкова по-малко часове спят нощем.

3. Заключение

След анализ на всички променливи стигнах до следните заключения:

- Най-голяма част от анкетираните са на възраст между 18 и 29 години.
- Предпочитаната напитка сред анкетираните консумиращи кофеин е кафе.
- Най-голяма част от приемащите кофеин анкетирани, консумират по 1 кофеинов продукт дневно.
- Има разлика в часовете сън между хората консумиращи кофеин и тези които не консумират, като тези които не консумират спят повече часове.
- Анкетираните, консумиращи повече кофеинови продукти дневно спят по-малко часове, в сравнение с тези, които консумират по-малко такива продукти.

Хипервръзка към dataset-а:
https://docs.google.com/spreadsheets/d/1Dl7K6p62SQ_bNP5X2bxOHtvxClBUkBGJ4cfnO54gzo4/edit?usp=sharing