Statistical analysis

Rubén Sánchez Fernández

09, octubre, 2018

Contents

This project is created as an exercise for *Analytical software for statistical analysis* course of Bioinformatics and Biostatistics MSc. (UOC)

Statistical analysis

In this report, we will perform a statistical analysis to a biomedical dataset.

The dataset

The data belongs to the *Duke University Cardiovascular Disease Databank* and includes records and features of patients under diagnostic cardiac catheterization. The dataset has 3504 patients and 6 features.

Data obtained from http://biostat.mc.vanderbilt.edu/DataSets

Data overview

We will start by importing the data.

```
library(Hmisc)
```

```
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
getHdata(acath)
```

Let's check we imported the righ dataset.

```
dim(acath)
```

```
## [1] 3504    6
```

The dataset has 3504 samples and 6 features, as we introduced in the previous section.

Let's use the *summary*, *str* and *names* functions to overview the data.

```
names(acath)
```

```
## [1] "sex"      "age"      "cad.dur"  "choleste" "sigdz"    "tvdlm"
```

```
summary(acath)
```

```
##       sex              age           cad.dur       choleste
##  Min.   :0.0000   Min.   :17.00   Min.   :  0   Min.   : 29.0
##  1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:  4   1st Qu.:196.0
##  Median :0.0000   Median :52.00   Median : 18   Median :224.5
##  Mean   :0.3136   Mean   :52.28   Mean   : 43   Mean   :229.9
##  3rd Qu.:1.0000   3rd Qu.:59.00   3rd Qu.: 60   3rd Qu.:259.0
##  Max.   :1.0000   Max.   :82.00   Max.   :416   Max.   :576.0
##                                                 NA's   :1246
##      sigdz            tvdlm
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
##  Mean   :0.6661   Mean   :0.3225
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##                   NA's   :3
```

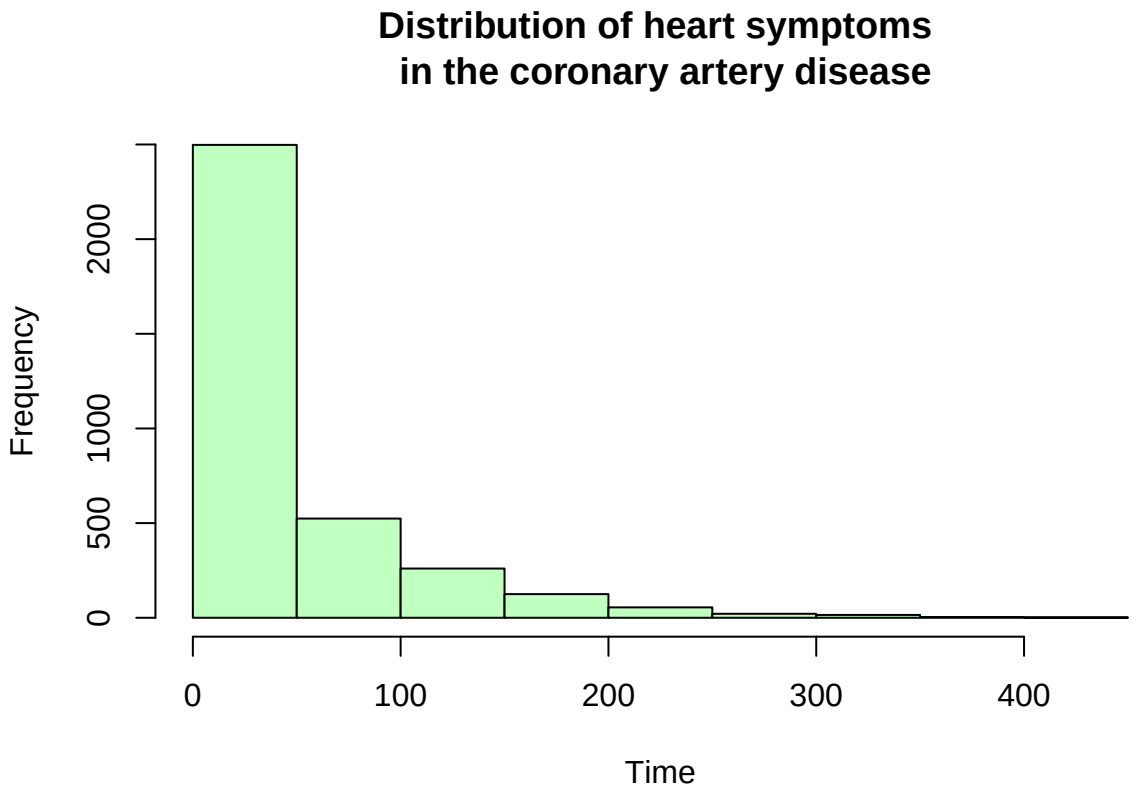
```
str(acath)
```

```
## 'data.frame':    3504 obs. of  6 variables:
##  $ sex     : 'labelled' int  0 0 0 1 1 0 0 0 0 0 ...
##  $ age     : 'labelled' int  73 68 54 58 56 64 65 41 68 52 ...
##   ..- attr(*, "label")= chr "Age"
##   ..- attr(*, "units")= chr "Year"
##  $ cad.dur : 'labelled' int  132 85 45 86 7 0 76 15 30 1 ...
##   ..- attr(*, "label")= chr "Duration of Symptoms of Coronary Artery Disease"
##  $ choleste: 'labelled' int  268 120 NA 245 269 NA NA 247 NA NA ...
##   ..- attr(*, "label")= chr "Cholesterol"
##   ..- attr(*, "units")= chr "mg %"
##  $ sigdz   : 'labelled' int  1 1 1 0 0 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Significant Coronary Disease by Cardiac Cath"
##  $ tvdlm   : 'labelled' int  1 1 0 0 0 0 1 0 1 0 ...
##   ..- attr(*, "label")= chr "Three Vessel or Left Main Disease by Cardiac Cath"
##  - attr(*, "comment")= chr "Data from the Duke University Cardiovascular Disease Databank"
```

From the overview, we can notice that we have several *missing values*, specially in *choleste* and *tvdlm* features. Let's study the features more closely.

For instance, let's plot the distribution of *cad.dur* variable, which is the duration of the heart symptoms.

```r
hist(acath$cad.dur, col="darkseagreen1", main="Distribution of heart symptoms \n in the coronary artery
```



We could also check the maximum and minimum of the variable *age*.

```r
print(paste0("The older patient is ", max(acath$age), " years old"))
```

```
## [1] "The older patient is 82 years old"
```

```r
print(paste0("and the youngest patient is ", min(acath$age), " years old"))
```

```
## [1] "and the youngest patient is 17 years old"
```

Another interesting question we could make is how many patients had high level of cholesterol (more than 200).

```r
nrow(subset(acath, acath$choleste>200))
```

```
## [1] 1602
```

1602 patients had high cholesterol.

Probability and simulation

Now, let's use probaility theory to answer some questions:

- What is the probability to find more than 60 patients with the disease if 100 tests were performed?

This case follows a binomial distribution, therefore we need to find the probability to have the disease from the data.

```
est<-subset(acath,acath$sigdz==1) #patients with the disease
p<-nrow(est)/nrow(acath) #probability = patients with the disease/total patients
```

Once we have the probability p, we can solve the question:

```
p60<-1-pbinom(60,100,p,lower.tail=T) #dist binomial
print(paste0("The probability to find more than 60 patients with the disease after evaluating 100 tests
```

```
## [1] "The probability to find more than 60 patients with the disease after evaluating 100 tests is 0.9
```

- Probability of not having the disease if someone has a narrow coronary artery?

This follows conditional probability as:

$$P(no\ disease|narrow\ artery) = \frac{P(no\ disease \bigcap narrow\ artery)}{P(narrow\ artery)}$$

Let's calculate the probabilities:

```
sya<-subset(acath,acath$tvdlm==0 & acath$sigdz==1)
psya<-nrow(sya)/nrow(acath) #probability no disease and narrow artery
ps<-psya/p #conditional probability
print(paste0("Probability of not having the disease while having narrow coronary artery is ",ps))
```

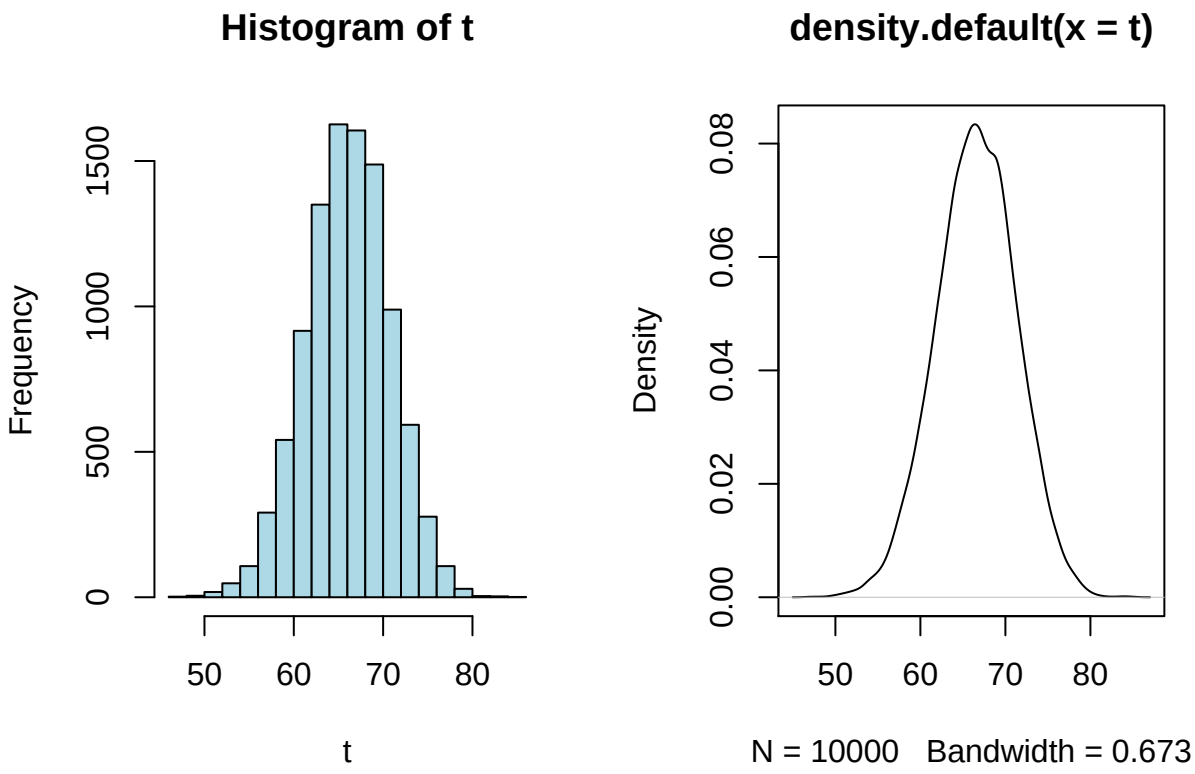
```
## [1] "Probability of not having the disease while having narrow coronary artery is 0.515424164524422"
```

In some situations, performing simulations can be very useful when analyzing data. For instance, let's simulate 10,000 times the number of patients with a narrow coronary artery if 100 tests were performed.

```
set.seed(12345)
t<-rbinom(10000,100,p)
str(t)
```

```
##  int [1:10000] 72 65 67 63 74 63 71 72 66 61 ...
```

```
#plots
par(mfrow=c(1,2))
hist(t,col="lightblue")
plot(density(t))
```




```
mean(t)
```

```
## [1] 66.5515
```

```
sd(t)
```

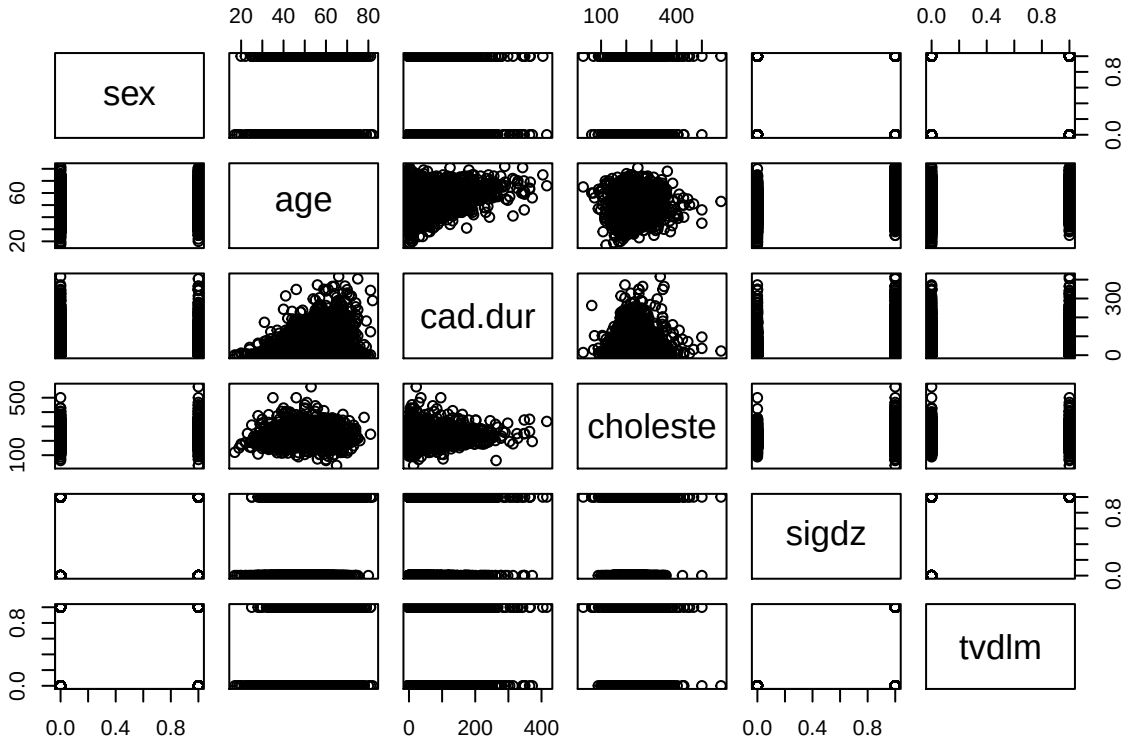
```
## [1] 4.718429
```

As we could expect, the mean obtained simulating 10,000 times is very close to the 1-time mean. It is interesting that the standard deviation is not important, therefore there is not much variability in the process.

Regression analysis

Regression analysis is also very useful when answering questions about our data. For instance, let's question how age influences in the duration of the symptoms.

```
pairs(acath)
```



Let's generate the linear regression model with *symptoms duration* and *age*.

```
RegModel <- lm(cad.dur~age, data=acath)
summary(RegModel)
```

```
##
## Call:
## lm(formula = cad.dur ~ age, data = acath)
##
## Residuals:
## Duration of Symptoms of Coronary Artery Disease
##    Min     1Q Median     3Q    Max
## -89.56 -34.89 -16.10  17.07 349.95
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.79200    5.04381  -8.881   <2e-16 ***
## age           1.67937    0.09479  17.717   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.72 on 3502 degrees of freedom
## Multiple R-squared:  0.08226,    Adjusted R-squared:  0.082
## F-statistic: 313.9 on 1 and 3502 DF,  p-value: < 2.2e-16
```

Determination coefficient is 0.08226. The model has a low explanatory power, so the model is not idoneal or the variables are independent.

Next, let's calculate the correlation coefficient between these variables.

```
cor(acath[,c("age","cad.dur")])
```

```
##               age  cad.dur
## age     1.000000 0.286806
## cad.dur 0.286806 1.000000
```

The coefficient is 0.2868. As we suspected, the dependency is low.

After this analysis, we can conclude that variable *age* doesn't influence in a significant level to the variable *symptoms duration*.