Introduction to R for Basic Statistics

Alessandra Meddis

Section of Biostatistics, University of Copenhagen

Day 3: 23 January, 2025

Outline

- Basic Graphs with `R`
  - Scatterplot
  - Boxplot
  - Histogram
- Exercise : descriptive analysis + plot

# Basic Graphs with R

General specification for graphs in R

For all graphics (plots, histograms,boxplots)

- col: color(s) of what you are drawing
- xlim,ylim: margins of the plot
- xlab,ylab: labels of the axes
- main: title of the plot
- lty: type of lines
- pch: type of points
- lwd: line width
- size: point size

Scatterplot

```
plot(db1_ex$height,db1_ex$weight,
     ylab='weight', xlab='height', ylim=c(50,70),
     col='red', pch=15,  main='Scatterplot')
```
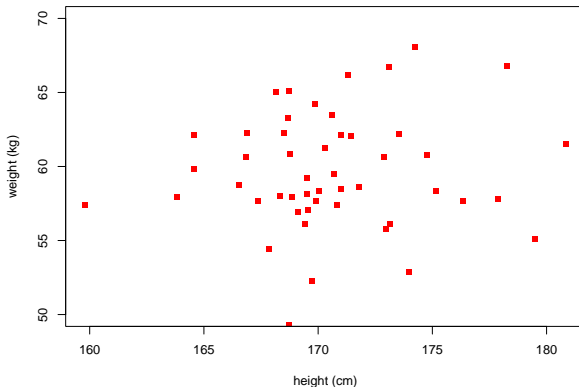
Scatterplot

- Use the function `plot()`. This will print points at the given coordinates.
- If only one coordinate is provided, then the variable is shown on the y-axis.
- If two coordinates are provided, the first argument is on the x-axis, and the second one on the y-axis.
- We can add lines and points on the top of the plot using the functions `points(),lines(),ablines()`

Add lines: abline()

```
abline(v=mean.height, col="green", lty=2)
abline(h=mean.weight, col="blue", lty=2)
points(mean.height,mean.weight, lwd=5)
```
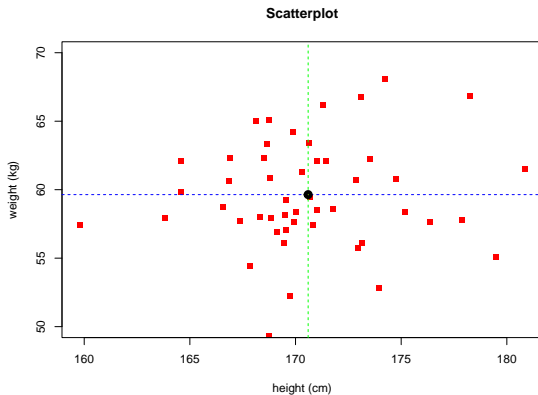
Add lines: lines()

```
plot(db1_ex$height,db1_ex$weight, ...,col=db1_ex$year)
abline(lm(weight~height, db1),lty=1, lwd=2)
legend( 'topleft', levels(db1_ex$years), pch=15, col=1:2,
    title='Year' )
```
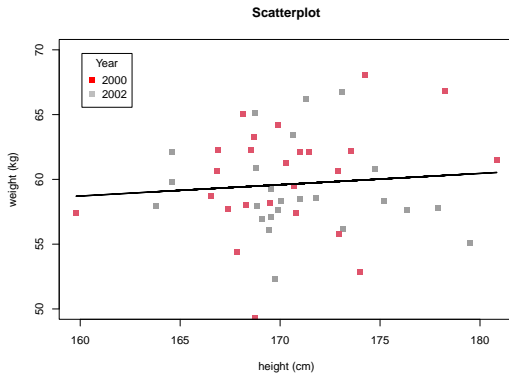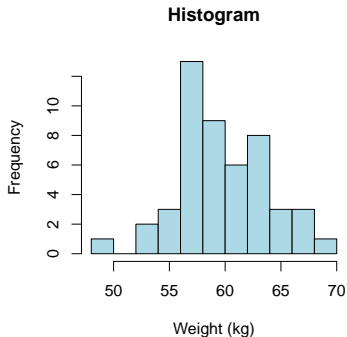
Add colors by a grouping variable

We specify the color with the argument `col`.

- The grouping variable defining the color needs to be a factor
- In the legend, to ensure that we are using the colors coresponding to the plot:

  - `levels()` : provides all possible levels for the grouping variable
  - `col= 1:nlevels(variable)`

- If we want to specify the colors, we need to add a column in the data frame, or define a vector where per each element we have the color
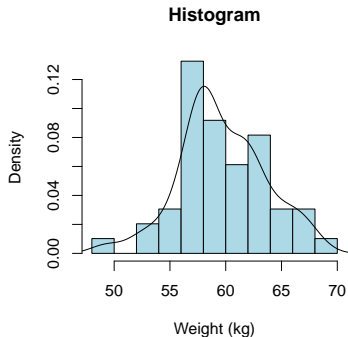
  ```
  db_long$color<-ifelse(db_long$year=="2000","red","blue")
  ```

Histogram

```
hist( db1_ex$weight, col="lightblue",     hist( ..., prob = TRUE)
xlab="Weight (kg)", main="Histogram")     lines(density(db1_ex$weight) )
```


Histogram

Histogram

Boxplot

A boxplot is used to illustrate key features of the distribution of a numerical variable for different groups.

```
boxplot(weight~year, db1_ex,
xlab='Year',names=c('2000','2002'),col=c("gray","red"),
```
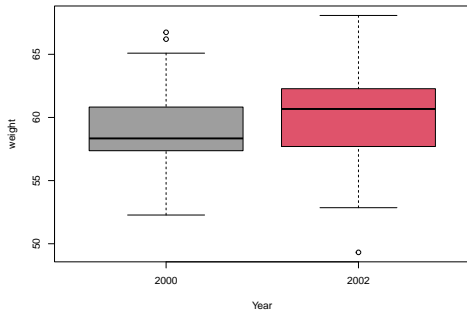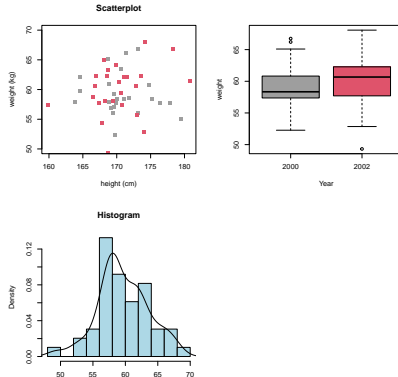
par()

The function par() is used to set several graphical parameters. This can be used to combine together several plots with the argument $= c(nrow, ncol)$

```
par(mfrow = c(2, 2))
# divide the space in 2 rows and 2 columns--> 4 plots max
```

General specification for plot

- `plot()`: Scatterplot or lines (check `type`)
- `points()`,`lines()`,`abline()`: to ADD points/lines to the plot
- For all graphics (plots, histograms,boxplots)
    - col: color(s) of what you are drawing
    - xlim,ylim: margins of the plot
    - xlab,ylab: labels of the axes
    - main: title of the plot
    - lty,pch: type of lines and points
    - lwd: line width

Exercise III

Load the data saved during the exercise on Wednesday or read the data in:
https://raw.githubusercontent.com/AMeddis/IntrotoR-for-Basic-Statistics/refs/heads/main/data_exercise/db_join.csv

1. *Make a boxplot of the diameter change by Day*

2. *Calculate the median change by treatment and Day (you can use the aggregate)*

3. *Rename the column "diam.change" with "median.change" of the data.frame obtained by the aggregate*

4. *Create a plot with the points at the median change from time 0 at varying of days where the color is defined by the treatment group. Add a legend.*