Scatter Diagrams

Intro to Stats, Spring 2017

Prof. Gaston Sanchez

Learning Objectives

- Becoming familiar with various measures of spread
- Intro to the functions `range()`, `IQR()`, and `sd()`
- Understand the concept of r.m.s. size of a list of numbers
- Be aware of the difference between SD and SD+

Introduction

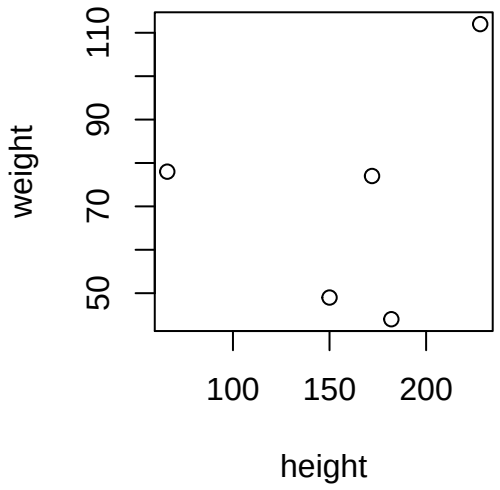
The easiest way to plot scatter diagrams in R is with the `plot()` function. I should say that `plot()` produces different kinds of plots depending on the type of input(s) that you pass to it.

If you pass two numeric variables (i.e. two R vectors) `x` and `y`, `plot()` will produce a scatter diagram.

Let's consider the following toy data table, and use `height` and `weight` to create a scatter diagram.

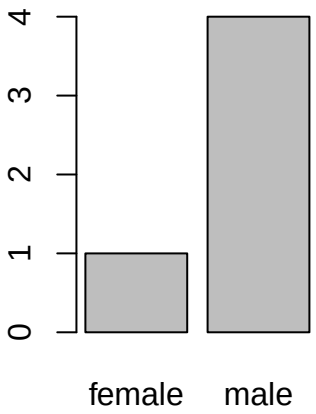
name	sex	height	weight
Luke	male	172.00	77.00
Leia	female	150.00	49.00
Obi-Wan	male	182.00	44.00
Yoda	male	66.00	78.00
Chebacca	male	228.00	112.00

```
height = c(172, 150, 182, 66, 228)  # in centimeters
weight = c(77, 49, 44, 78, 112)     # in kilograms

# default scatter diagram
plot(height, weight)
```



If you pass a factor to `plot()` it will produce a bar-chart:

```
# qualitative variable (as an R factor)
sex = factor(c('male', 'female', 'male', 'male', 'male'))

# default scatter diagram
plot(sex)
```

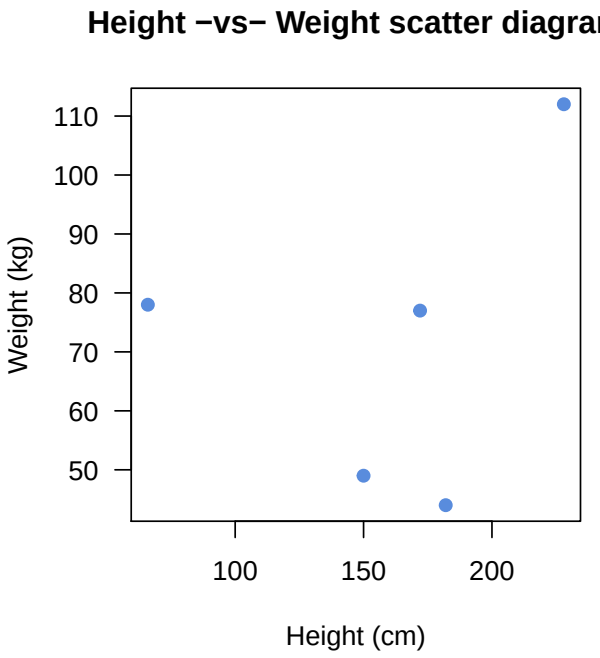


Note that `plot()` displays a very simple, and kind of ugly, scatter diagram. This not an accident. In fact, the basic plots in R follow a "quick and dirty" approach. They are not publication quality, but that is OK. The default display of `plot()` was not designed to produce pretty graphics, but rather to produce visualizations that quickly allow you to explore the data, identify patterns, help you ask new research questions, and then move on with more visualizations or to the next analytical stages.

Although `plot()` produces a basic graph, you can use several arguments, or graphical parameters, to obtain a nicer chart. To find more information about the available graphical parameters for `plot()`, take a look at the documentation provided by `help(plot)`.

The following code uses various graphical parameters to display a more visually appealing scatter diagram:

```
# nicer scatter diagram
plot(height, weight,
     las = 1,          # orientation of y-axis tick marks
     pch = 19,         # filled dots
     col = '#598CDD', # color of dots
     xlab = 'Height (cm)',    # x-axis label
     ylab = 'Weight (kg)',    # y-axis label
     main = 'Height -vs- Weight scatter diagram')
```



Scatter diagrams with `ggplot2`

Another approach to create scatter diagrams is to use functions from the package `"ggplot2"`. This package provides a different philosophy to define graphs, and it also produces plots with visual attributes carefully chosen to provide prettier plots.

You should have the package `"ggplot2"` already installed, since you were supposed to use it for HW02. Assuming that his is the case, you need to load `"ggplot2"` with the function `library()` in order to start using its functions:

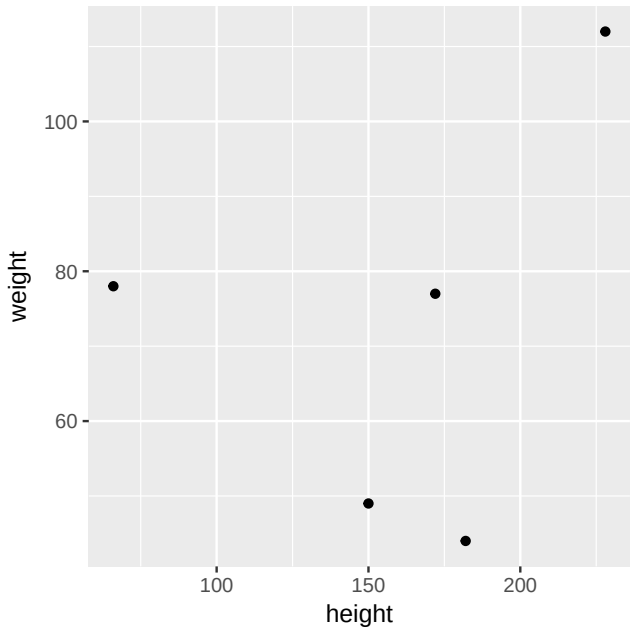
```
# load ggplot2
library(ggplot2)
```

One of the major differences between basic plots—like those produced by `plot()`—and graphics with `ggplot()`, is that the latter requires the data to be in the form of a data frame:

```
dat = data.frame(
  name = c('Luke', 'Leia', 'Obi-Wan', 'Yoda', 'Chebacca'),
  sex = c('male', 'female', 'male', 'male', 'male'),
  height = c(172, 150, 182, 66, 228),
  weight = c(77, 49, 44, 78, 112)
)
```

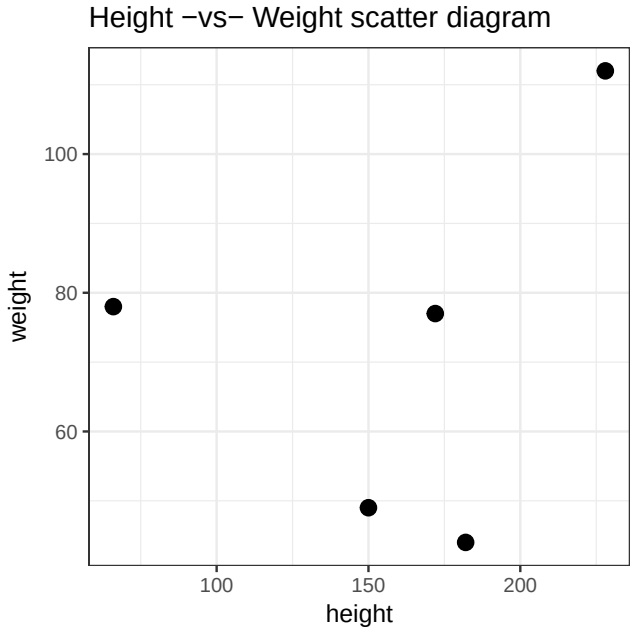
To create a scatter diagram with `"ggplot2"`, type the following commands:

```
ggplot(data = dat, aes(x = height, y = weight)) +
  geom_point()
```



As you can tell, the default chart produced by `ggplot()` is nicer than the one produced with `plot()`. You can customize the previous graph to add more details:

```
ggplot(data = dat, aes(x = height, y = weight)) +
  geom_point(size = 3) +
  theme_bw() +
  ggtitle("Height -vs- Weight scatter diagram")
```



Here's another example of a scatter diagram that includes labels for each dot:

```r
ggplot(data = dat, aes(x = height, y = weight)) +
  geom_point(size = 3) +
  geom_text(aes(label = name), hjust=0, vjust=0) +
  xlim(0, 300) +
  theme_bw() +
  ggtitle("Height -vs- Weight scatter diagram")
```