Principal Component Analysis

Part 1 - Descriptive statistics and correlation analysis

The goal of this course is to give you all the information you need to understand PCA, as well as give you all the tools you need to implement it. We will focus on the practical aspects of PCA without neglecting the mathematics. Each part ends with several exercises to do. Some of them are very easy and some are more difficult. I highly recommend checking all the demonstrations presented during the course, for 2 main reasons: to check for errors and to help you understand the objects manipulated. I truly believe that manipulating equations is very helpful in getting a good feel for why things are the way they are.

> *"Statistics is about reducing the amount of data."* **R. Fisher**

Reminder of basic statistics For a sample size of n, the mean is defined as the average value of each individual:

$$\overline{x} = \frac{\sum x_i}{n}$$

But you can easily find examples where 2 different samples have the same mean but don't have the same distribution. One way to illustrate this, is to compute other statistic, like standard deviation, defined as:

$$\sigma_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

You can interpret it as (*more or less*), the average difference between an individual and the mean of the sample he comes from.

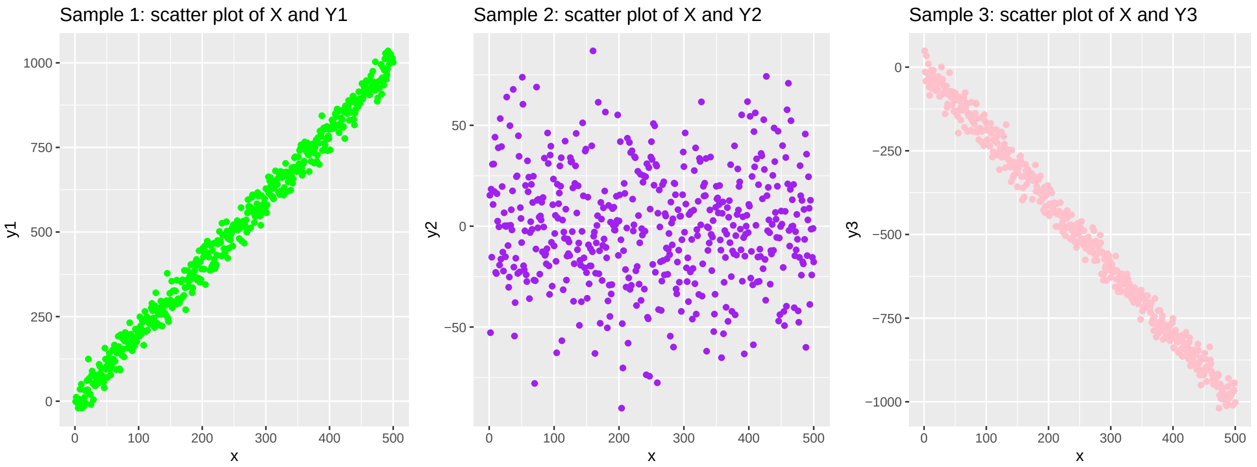
It's pretty common to also compute the variance, the squared standard deviation:

$$Var(x) = \frac{\sum (x_i - \overline{x})^2}{n}$$

These are the simplest univariate descriptive statistics you can find. You should be very comfortable working with them and interpret their value.

What is correlation? We want to know if 2 variables are correlated. But, if we ask ourselves what this formally means, it's not that easy to define. We will use the following definition: 2 variables are correlated if they tends to not be statistically dependent upon each other. Let's see a concrete example.

```r
#create different variables and check their correlation
n = 500
x = seq(1, n)
y1 = 2*x + rnorm(n=n, sd=30)
y2 = rnorm(n=n, sd=30)
y3 = -2*x + rnorm(n=n, sd=30)
data = data.frame(x, y1, y2, y3)
```



With plots, it's pretty obvious to detect when variables are correlated. So the question we want to ask is: **how can we *objectively* qualify correlation** (i.e: without using plots)?
For this, we use **Pearson correlation coefficient**:

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

The function $cov(x, y)$ is the covariance between x and y. The latter tends to tell us if, in a finite sample, when x_i is above (or below) the mean of its sample, y_i tends to also be above (or below) the mean of its sample. **The higher the covariance is, the more when x_i is above its mean, y_i also is**. To the contrary, the lower the covariance is, the more when x_i is above (resp.below) its mean, y_i is below (resp. above). More formally, we can describe it this way:

$$cov(x, y) = \frac{1}{n} \sum (x_i - \overline{x})(y_i - \overline{y})$$

Covariance actually measures correlation between 2 variables. But **Pearson correlation coefficient is better because it standardizes the covariance between -1 and 1**. In fact, covariance unit is the product of the units of x and y: doesn't make lot of sense for us. That's why we divide it by the product of standard deviation.

Pearson correlation coefficient can be interpreted as follow: **the closer it is to 1 (resp. -1), the more there is a positive (resp. negative) correlation**. If it's near 0, it seems that there is no correlation between variables (*that is not completely true, but we will do like it is*). If we take back our last sample back, we can compare their correlation coefficient.
Ask yourself before checking the results: which one will be the highest? Why?

```
## Correlation between variables from our first sample: r = 0.9950904
```

```
## Correlation between variables from our second sample: r = -0.0327109
```

```
## Correlation between variables from our third sample: r = -0.9952475
```

The variables from sample 1 and 3 are very correlated (r is near 1 and -1).
The 2 other variables are much less correlated (r is near 0).

Interlude: non-linear correlation

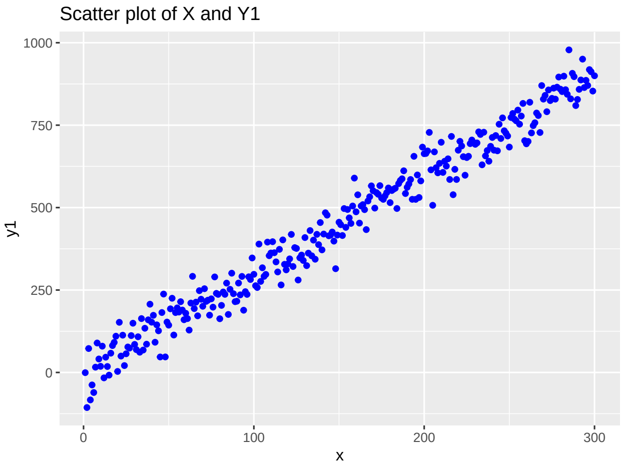
Since the Pearson coefficient correlation only measures monotonic linear correlation, **it will miss correlation that are not linear (but still monotonic)** like exponential or logarithmic. In order to solve this

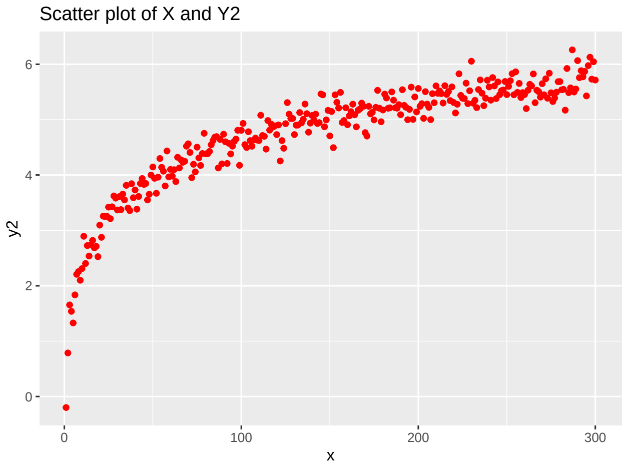
issue, statisticians have invented other measure of correlation. We will discuss the main one: **Spearman correlation coefficient**.
It's defined as follow:

$$s = \frac{\frac{1}{n}\sum(rx_i - \overline{rx}_n)(ry_i - \overline{ry}_n)}{r_\sigma(x)r_\sigma(y)}$$

Where rx_i design the rank of the i-th individual for the variable x compared to other individuals. Using this statistic compared to the Pearson's one allow us to measure non-linear relationship. Spearman's coefficient is considered as **non-parametric** because it doesn't use the value of x_i but its rank. Let's see an example of when this might be useful.

```
#create different variables and check their correlation
n = 300
x = seq(1, n)
y1 = 3*x + rnorm(n=n, sd=45)
y2 = log(x) + rnorm(n=n, sd=0.2)
data = data.frame(x, y1, y2)
```





Compute both Pearson and Spearman coefficients correlation

```
## Pearson correlation between variables from our first sample: r = 0.9860032
## Spearman correlation between variables from our first sample: r = 0.987007

## Pearson correlation between variables from our second sample: r = 0.8647263
## Spearman correlation between variables from our second sample: r = 0.9450163
```

Spearman coefficient is higher for the sample with a logarithmic (non-linear) relationship. This tool is useful to illustrate that 2 variables can be considered as correlated but their relationship is more complex that a straight line.

Statistical significance of correlation

For the moment, we only said that the closer r was to 1 or -1, the stronger the correlation was. But how can we know if this is due to luck or a real correlation? To answer this question we can compute a Student's test on r. This tells us, assuming that there is no correlation between the variables, **what was the probability (named p-value or p) that we actually observe the current value of r.** If the latter is too low (convention is 5%), we consider that our assumption might be false and that there is actually a correlation. Let's see how it works with R.

```
data("iris")
cor.test(iris$Sepal.Width, iris$Petal.Length)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Width and iris$Petal.Length
## t = -5.7684, df = 148, p-value = 4.513e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5508771 -0.2879499
## sample estimates:
##        cor
## -0.4284401
```

In this case, $r = -0.428$ and $p < 0.05$. We then consider that there is a significant negative correlation between Sepal.Width and Petal.Length.

Why are we interested in correlation? Depending of what you're interested in (descriptive or inferential statistics), you will not use correlation for the same reason and the same way. **Correlation is an important statistic tool used in a daily basis in lots of field**. For example, if we observe a strong correlation between a genome and breast cancer, we might do some prevention for people with this genome and easily attenuate the cancer impact, especially if it's not already present.
However, **you should always have in mind that correlation has not so much to do with causality**. The latter is way more complex and will not be discussed here.