

# Initiation to R Software

*Pierre Michel*

*Master AMSE 1st year, 2019*

## Problem Set III

### 1) Import and export data

- Create an ASCII file (a text file with tabulation separator) with 4 columns (3 **numeric**, 1 **character**) and 5 observations. Give the columns names in the first row. Save this file in the working directory.
- Import this file in R, in an object of type **data.frame**. Print the content of the object, check rows/columns names (**name()**, **dimnames()**, **colnames()**, **rownames()**).
- Choose rows names.
- Add a column whose values indicate if values in the third columns are greater than 10 (1 if  $> 10$ , 0 elsewhere).
- Export the new **data.frame** to the working directory in a text file, with space separator, without rows/columns names.

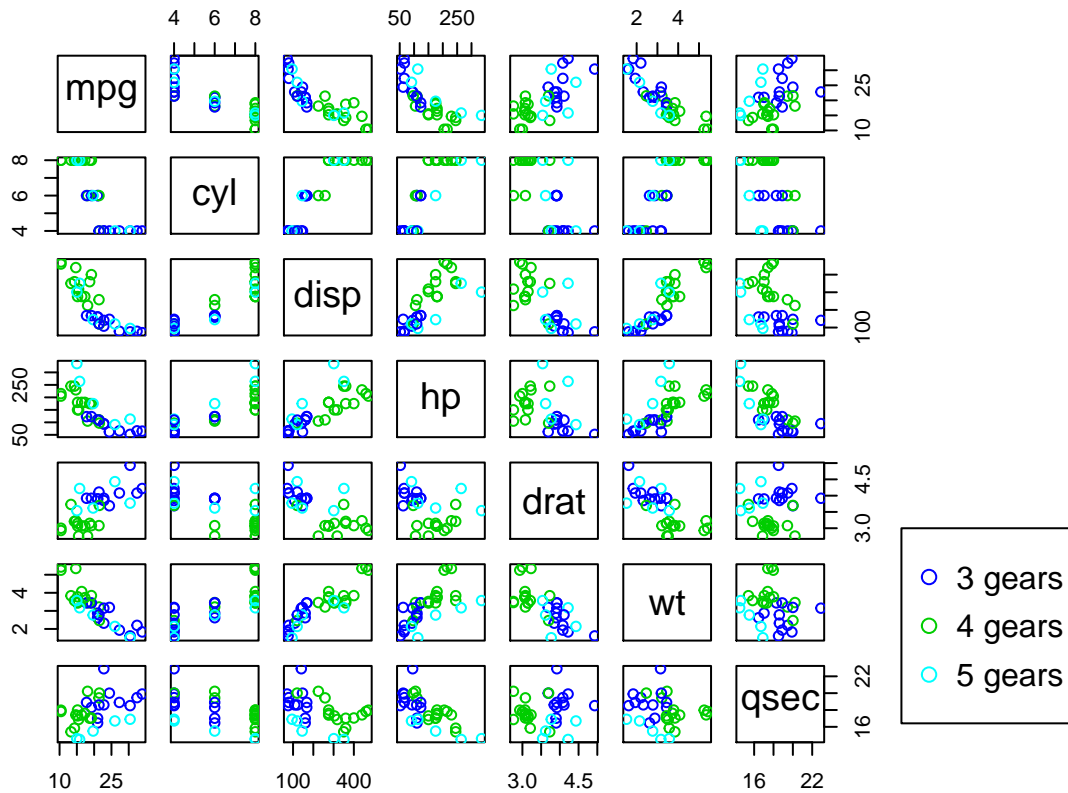
### 2) Probability distributions

- Compute the probability that a random variable  $X$  following a binomial distribution  $\mathcal{B}(10, \frac{1}{3})$  takes the value 0, 1, 2, ..., 10, i.e  $P(X = k)$ .
- Compute the probability that a random variable  $X$  following a binomial distribution  $\mathcal{B}(10, \frac{1}{3})$  takes values less than or equal to 10 and greater than 5, i.e  $P(5 < X \leq 10)$ .
- Compute the value  $x$  for which  $P(X \leq x) = 0.97$ , where  $X$  follows a normal distribution  $\mathcal{N}(0, 1)$ .
- Compute the quantile of order 2% of a Student's distribution with 5 degrees of freedom ( $\sqcup(5)$ ).

### 3) Descriptive statistics

Load the *mtcars* dataset from the package **base**, and print its content. Below is a statistical summary of the dataset and a plot of pairs of continuous variables.

```
##      mpg      cyl      disp      hp
## Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
## 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
## Median :19.20   Median :6.000   Median :196.3   Median :123.0
## Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
## 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
## Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##      drat      wt      qsec      vs
## Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
## 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
## Median :3.695   Median :3.325   Median :17.71   Median :0.0000
## Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
## 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
## Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##      am      gear      carb
## Min.   :0.0000   Min.   :3.000   Min.   :1.000
## 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
## Median :0.0000   Median :4.000   Median :2.000
## Mean   :0.4062   Mean   :3.688   Mean   :2.812
## 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
## Max.   :1.0000   Max.   :5.000   Max.   :8.000
```



- For continuous variables you can have a statistical summary with standard empirical estimations: mean, variance, first and third quartiles and median. Compute these *mtcars* columns that correspond to continuous variables (use `summary()`). Plot the histograms/boxplots (use `hist()` and `boxplot()`).
- Identify the variables for which the assumption of normality is plausible (use `qqnorm()` and `qqline()`).
- Evaluate the possible correlations between these variables (use `cor()`).
- For discrete variables, provide a table of counts (use `table()`) and a graphical representation of this table (`barplot(table())`). Comment.

#### 4) Generate random samples, empirical mean and variance

- Generate 100 samples of size 100 from a normal distribution  $\mathcal{N}(3,1)$ . For each sample, compute the mean and the empirical variance (use `apply()`, `replicate()`).
- Compute the mean and variance of the two series.
- Plot on the same figure the histogram of relative counts with a kernel estimation of the means series.
- Re-run for 100, 500, 1000 samples. What do you observe ? Comment the results.

#### 5) Random sampling

- Simulate 25 draws of a coin.
- Test the following commands:
  - `urn = c(rep("red", 8), rep("blue", 4), rep("yellow", 3)); sample(urn, 6, replace = F)`
  - `plot(0:10, dbinom(0:10, size = 10, prob = .25), type="h", lwd = 30, col = "gray", main="Binomial distribution; n=10; p=0.25")`
  - `curve(dnorm(x), from = -3, to = 3)`
  - `curve(pnorm(x, mean = 10, sd = 2), from = 4, to = 16)`