STT 1000 - STATISTIQVES

ARThVR ChARPENTIER

“Logiciels de Statistique”

cf comparaison des logiciels de statistiques sur wikipedia

La connaissance d’un logiciel de programmation/calculs/analyse de données est devenu un éléement incontournable pour les chercheurs, praticiens, industriels et donc des étudiants.

Pourquoi R ? Notamment parce qu’il est gratuit, libre, multiplateforme et bénéficie de la plus grande communauté au monde.

Langage S/Logiciel R/RStudio

S "*est un langage de programmation de très haut niveau et un environnement d'analyse des données et des graphiques conçu dans les années 1975-1976 par John Chambers. Les deux interpréteurs modernes de S sont S-Plus et R*"

R "*naît en 1993 comme un projet de recherche de Ross Ihaka et Robert Gentleman à l'université d'Auckland, sous une licence GNU GPL*" (logiciel libre et gratuit)

RStudio: Logiciel libre, multi-plateforme, gratuit ; Interface graphique permettant un environnement de développement intégré (IDE).

Un IDE n'est pas une interface graphique au sens de SPSS ou SAS, qui permettrait d'utiliser le logiciel à travers des menus et des boîtes de dialogue : il s'agit d'un environnement facilitant la saisie, l'exécution de code, la visualisation des résultats, etc.

Foire aux Questions

- *Ai-je besoin d'installer les logiciels R et RStudio sur une machine personnelle ?*

Oui, ça sera plus simple / **Non**, il existe de versions en ligne (https://rstudio.cloud/)

- *Ai-je besoin de maîtriser parfaitement le logiciel R pour réeussir le cours avec succès ?*

Non, la maîtrise de R n'est pas une fin en soi (dans ce cours). R est juste un outils (très pratique)

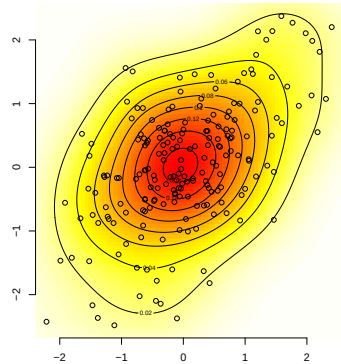
- *R/RStudio sera-t-il utilisé comme outil d'évaluation ?*

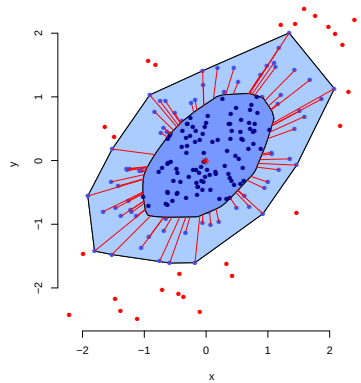
Oui, les examens pourront contenir des sorties logicielles qu'il conviendra d'interpréter,

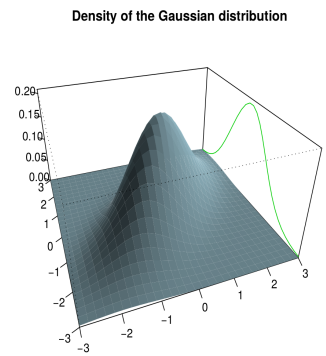
S/R Language

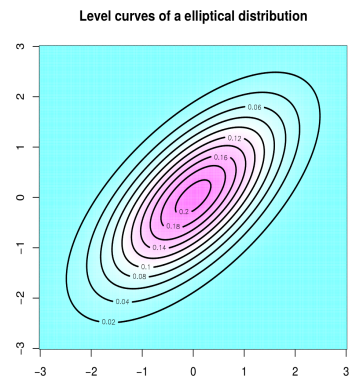
Statistiques et graphiques

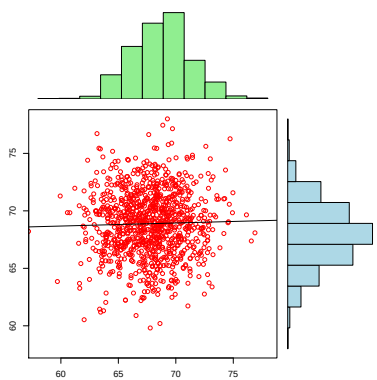
```
> x = c(0,1,2,3,4,5,6,7,8,9,10)
> mean(x)
[1] 5
> quantile(x, .95)
95%
9.5
> quantile(x, .95, type=1)
95%
 10
> quantile(x, .95, type=4)
 95%
9.45
> quantile(x, .95, type=5)
 95%
9.95
```

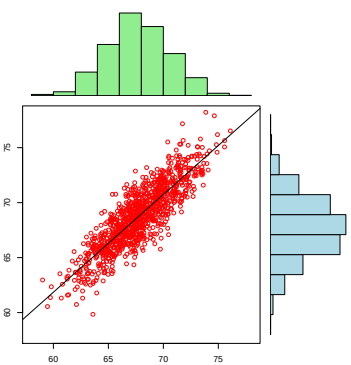




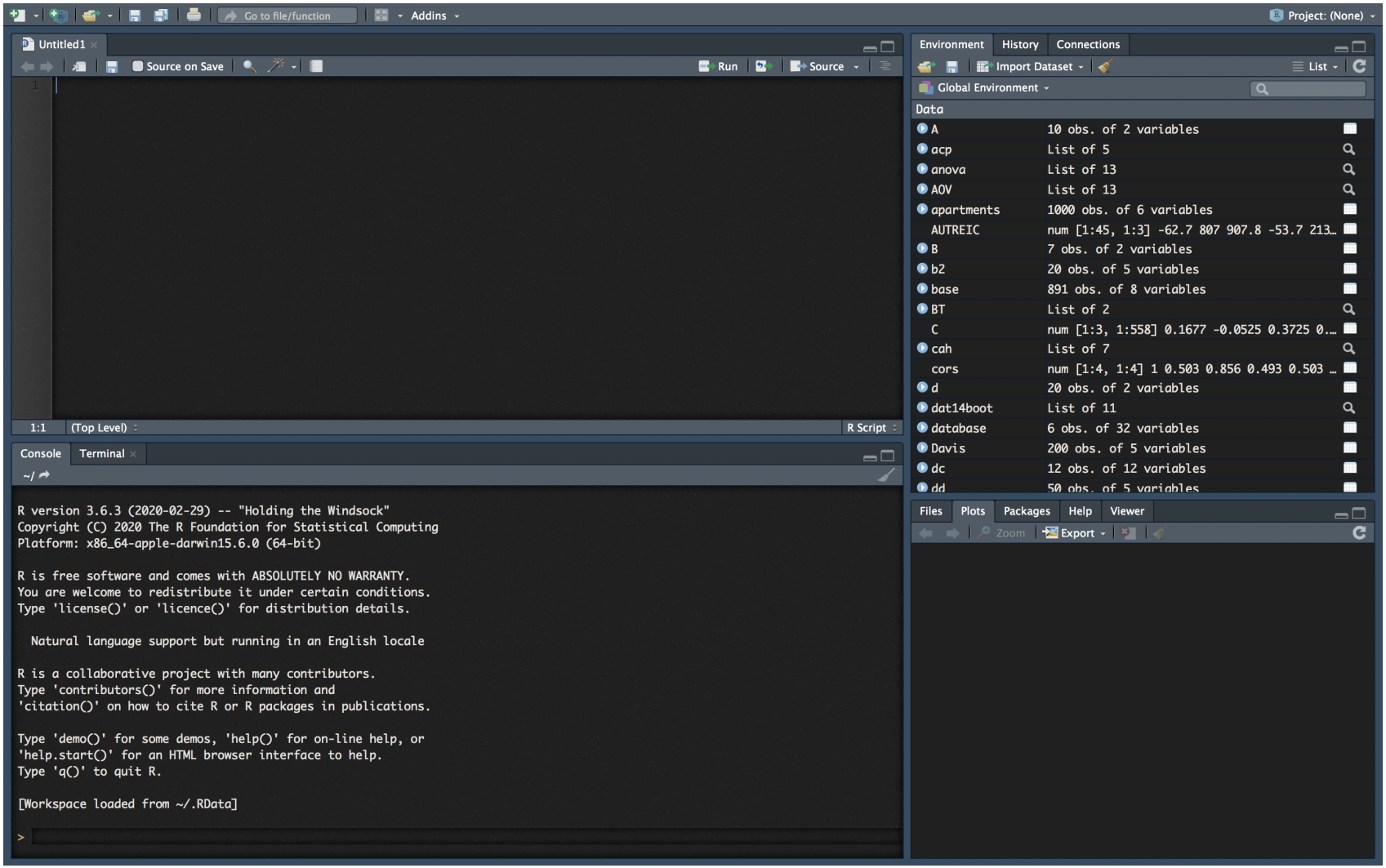






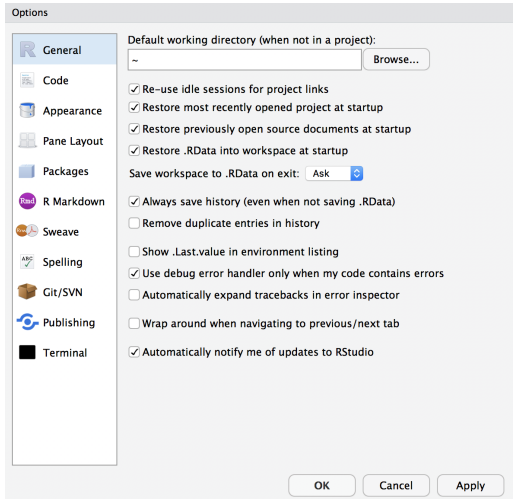


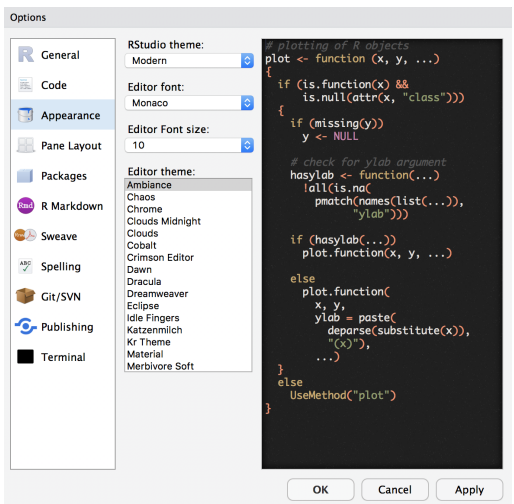
RStudio

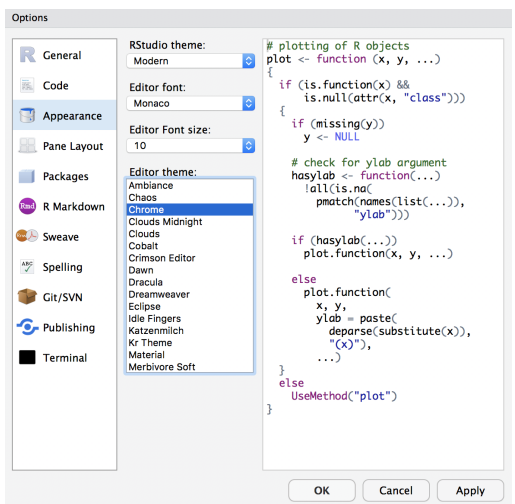

Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Console
Terminal
~/
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
A 10 obs. of 2 variables
acp List of 5
anova List of 13
AOV List of 13
apartments 1000 obs. of 6 variables
AUTREIC num [1:45, 1:3] -62.7 807 907.8 -53.7 213…
B 7 obs. of 2 variables
b2 20 obs. of 5 variables
base 891 obs. of 8 variables
BT List of 2
C num [1:3, 1:558] 0.1677 -0.0525 0.3725 0.…
cah List of 7
cors num [1:4, 1:4] 1 0.503 0.856 0.493 0.503 …
d 20 obs. of 2 variables
dat14boot List of 11
database 6 obs. of 32 variables
Davis 200 obs. of 5 variables
dc 12 obs. of 12 variables
dd 50 obs. of 5 variables
Files
Plots
Packages
Help
Viewer
Zoom
Export

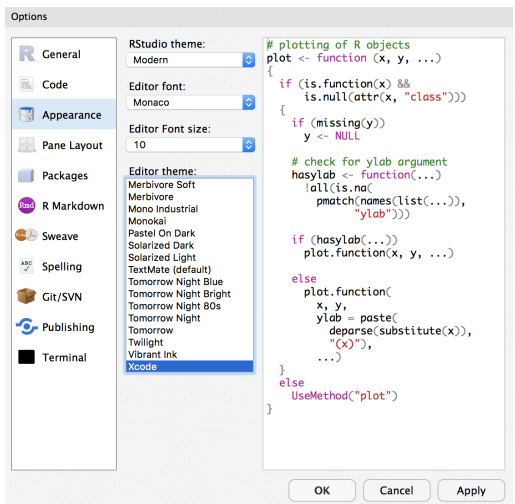
RStudio: appearance

Ma version de RStudio est sombre
on peut la paramétrer...

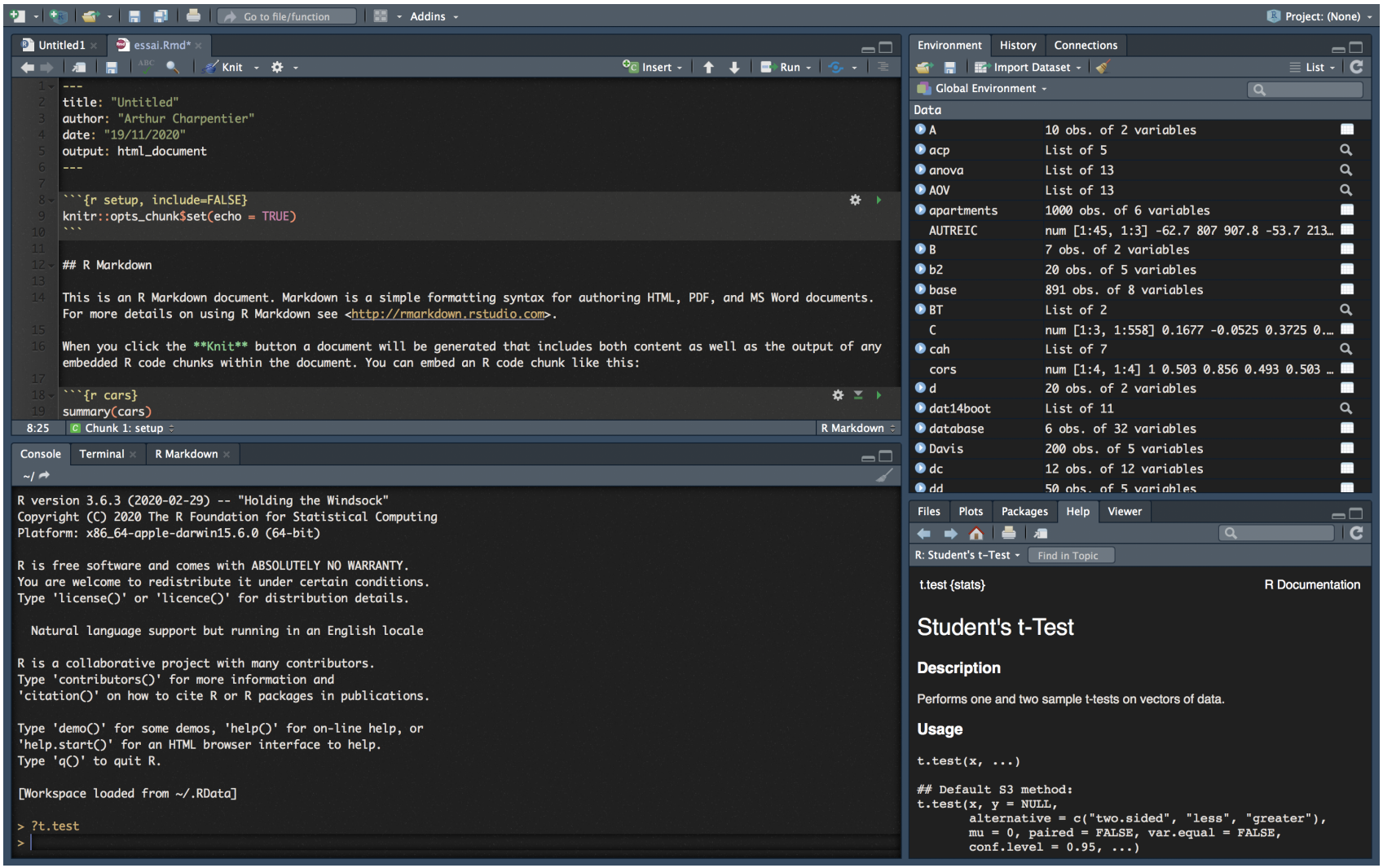






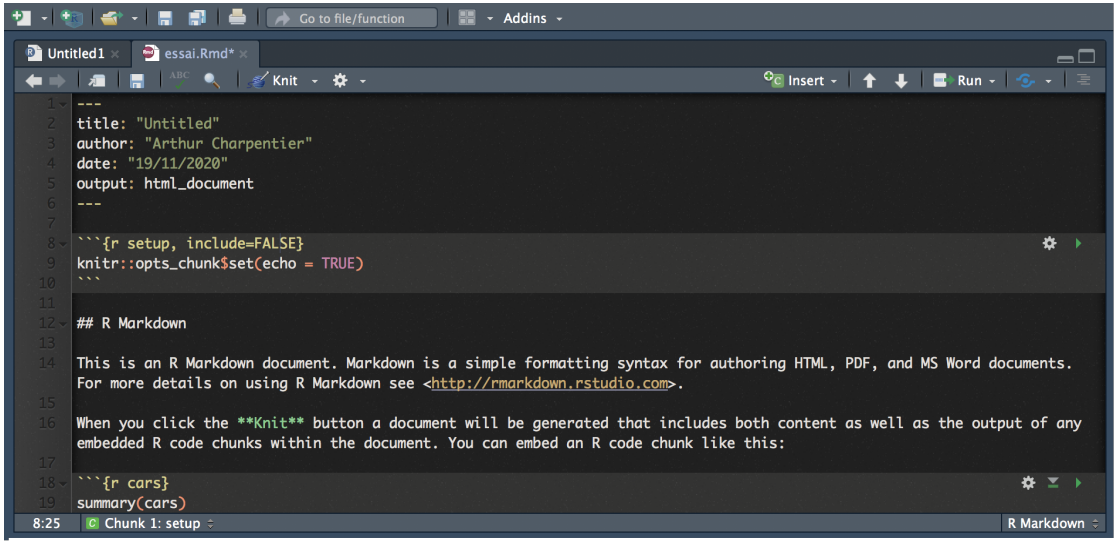


RStudio


Go to file/function
Addins
Project: (None)
Untitled1
essai.Rmd*
Knit
Insert
Run
---
title: "Untitled"
author: "Arthur Charpentier"
date: "19/11/2020"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.
For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any
embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
8:25
Chunk 1: setup
R Markdown
Console
Terminal
R Markdown
~/
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
> ?t.test
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
A 10 obs. of 2 variables
acp List of 5
anova List of 13
AOV List of 13
apartments 1000 obs. of 6 variables
AUTREIC num [1:45, 1:3] -62.7 807 907.8 -53.7 213…
B 7 obs. of 2 variables
b2 20 obs. of 5 variables
base 891 obs. of 8 variables
BT List of 2
C num [1:3, 1:558] 0.1677 -0.0525 0.3725 0.…
cah List of 7
cors num [1:4, 1:4] 1 0.503 0.856 0.493 0.503 …
d 20 obs. of 2 variables
dat14boot List of 11
database 6 obs. of 32 variables
Davis 200 obs. of 5 variables
dc 12 obs. of 12 variables
dd 50 obs. of 5 variables
Files
Plots
Packages
Help
Viewer
R: Student's t-Test
Find in Topic
t.test {stats}
R Documentation
Student's t-Test
Description
Performs one and two sample t-tests on vectors of data.
Usage
t.test(x, ...)
## Default S3 method:
t.test(x, y = NULL,
alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE,
conf.level = 0.95, ...)
```

RStudio



La fenêtre editor (haut à gauche)

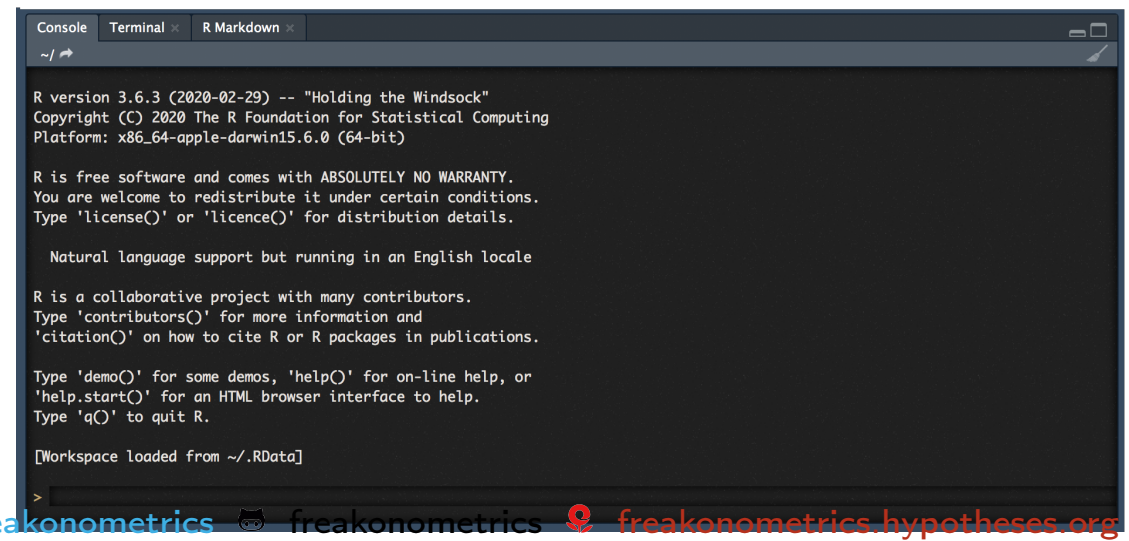
On peut y taper une succession de commandes (que l'on pourra exécuter ensuite)

```
# 5 observations
v = c(0.89367,-1.04729,1.97133,-0.38363,1.65414)
mean(v)
f = function(x) sum((v-x)^2)
optim(0, f)
```

RStudio

La fenêtre console (bas à gauche) : après avoir tapé '**run**',

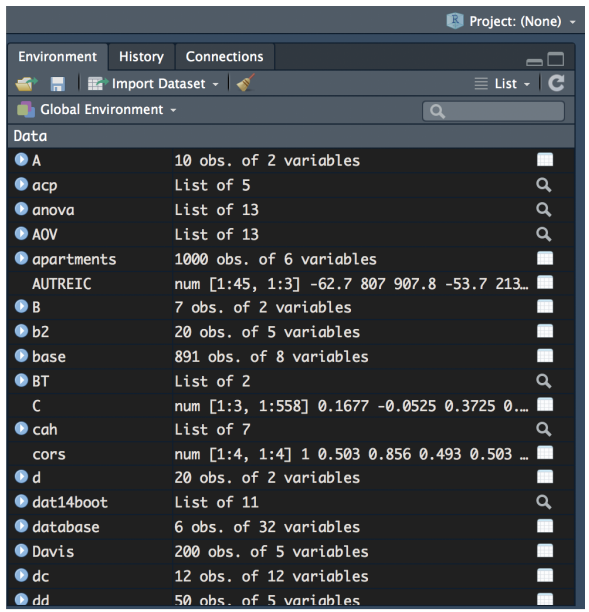
```
> v = c(0.89367,-1.04729,1.97133,-0.38363,1.65414)
> mean(v)
[1] 0.617644
> f = function(x) sum((v-x)^2)
> optim(0, f)
$par
[1] 0.6175781
Warning message:
In optim(0, f) : one-dimensional optimization by
    Nelder-Mead is unreliable:
use "Brent" or optimize() directly
```



RStudio

La fenêtre environment (haut à droite) donne la liste des objets en mémoire

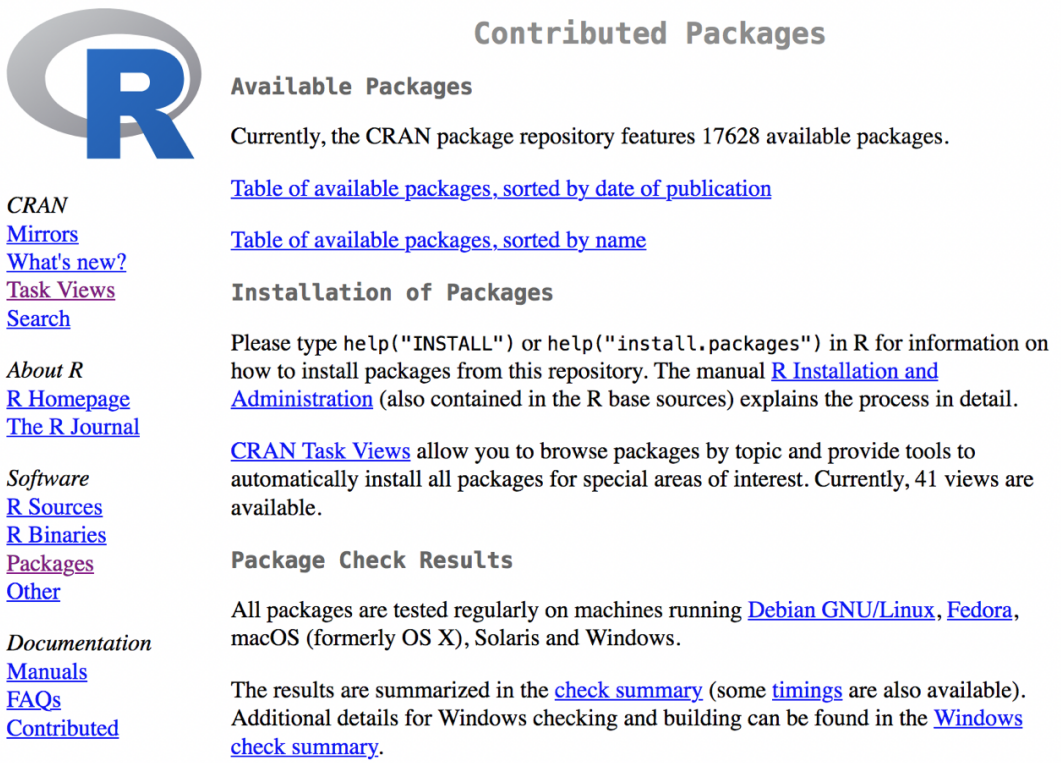
```
v     num [1:5] 0.894 -1.047...
f     function(x)
```



RStudio

La fenêtre packages (bas à droite)

```
1 > install.packages("vcd")
```



Contributed Packages

Available Packages

Currently, the CRAN package repository features 17628 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

Installation of Packages

Please type help("INSTALL") or help("install.packages") in R for information on how to install packages from this repository. The manual R Installation and Administration (also contained in the R base sources) explains the process in detail.

CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 41 views are available.

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux, Fedora, macOS (formerly OS X), Solaris and Windows.

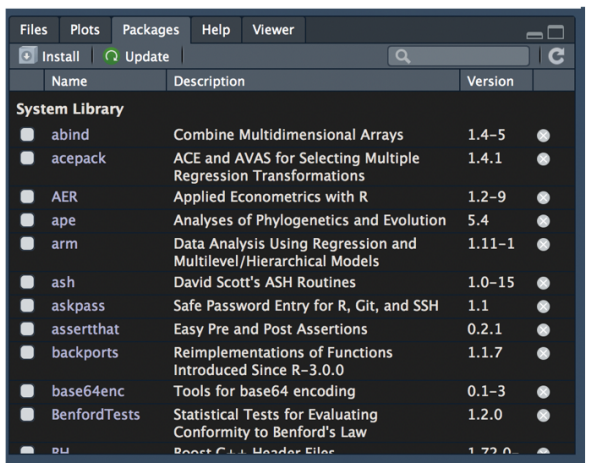
The results are summarized in the check summary (some timings are also available). Additional details for Windows checking and building can be found in the Windows check summary.

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed



RStudio: aide

Sample Quantiles

Description

The generic function quantile produces sample quantiles corresponding to the given probabilities. The smallest observation corresponds to a probability of 0 and the largest to a probability of 1

Usage

```
> quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
    names = TRUE, type = 7, ...)
```

Arguments

`x`: numeric vector whose sample quantiles are wanted, or an object of a class for which a method has been defined (see also 'details'). `NA` and `NaN` values are not allowed in numeric vectors unless `na.rm` is `TRUE`

References

Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, American Statistician 50, 361–365. doi: 10.2307/2684934.

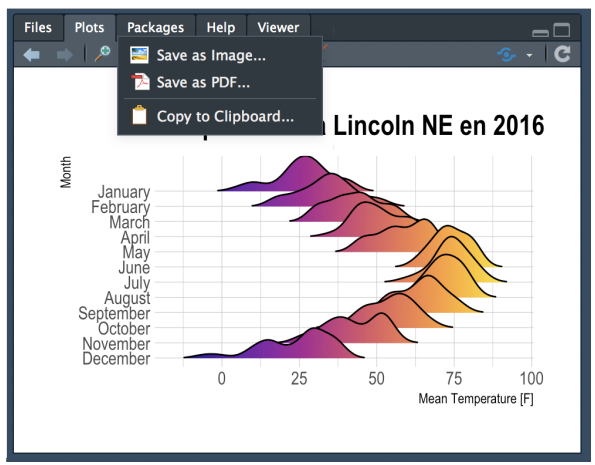
RStudio: graphiques

Pour les graphiques (de base)

1. il faut créer la fenêtre graphique (avec la fonction `plot`)

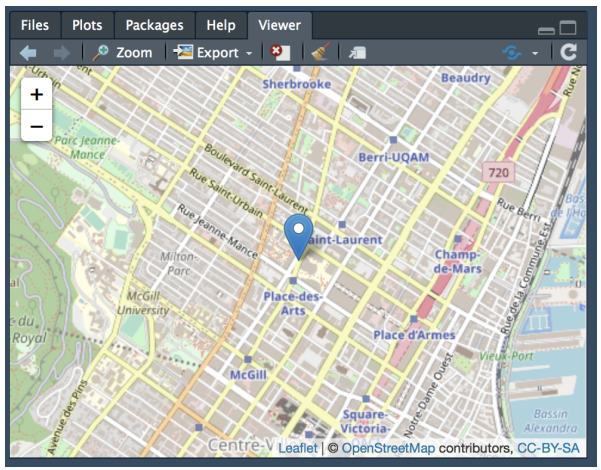
```
1 > plot(cars)
```

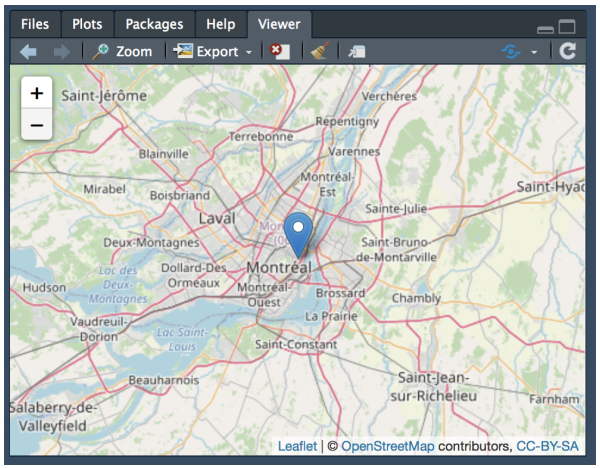
2. on peut alors superposer des courbes, des points, etc



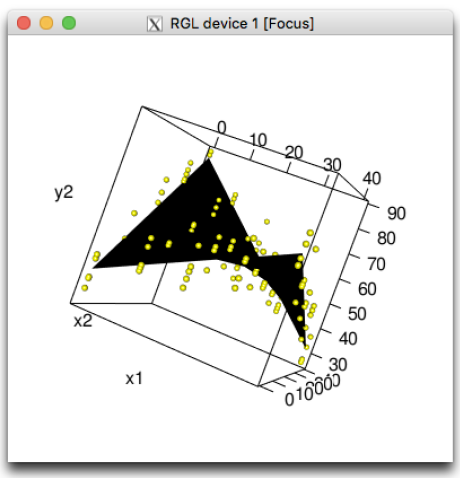
RStudio

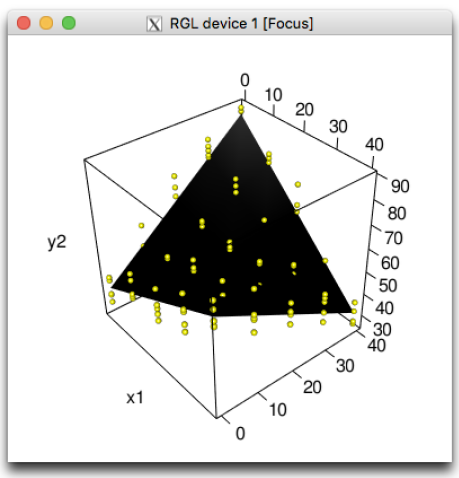
Certains graphiques (dynamiques) apparaîtront dans la fenêtre viewer (e.g. leaflet)

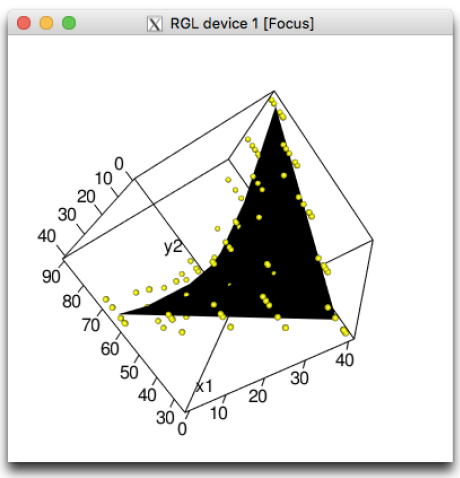




ou dans une fenêtre extérieure (rgl)



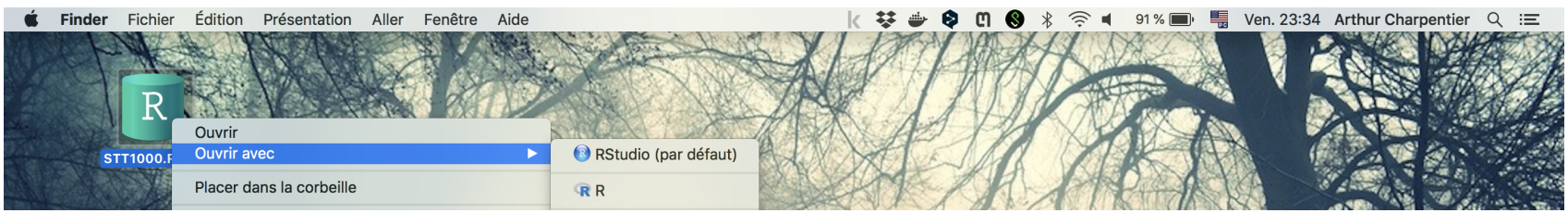




Fichier STT1000.RData

Les fichiers .Rdata contiennent des données (format lisible en R).

```
> url = "http://freakonometrics.free.fr/STT1000.RData"
> download.file(url,"STT1000.RData")
trying URL
Content type 'text/plain' length 459595 bytes (448 KB)
==================================================
downloaded 448 KB
> load("STT1000.RData")
```



Fichier STT1000.RData

```
> ls()
 [1] "a"                    "alcool"               "aspirine"
 [4] "attente"              "b"                    "babies"
```

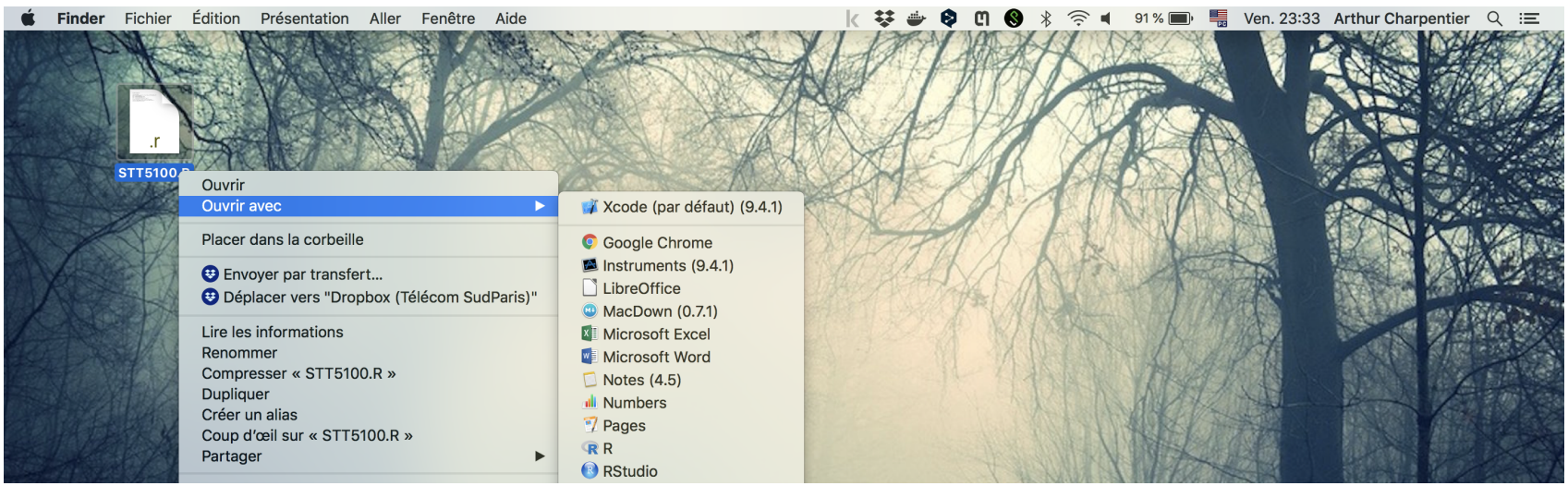
(etc)

```
> tail(Davis)
     sex weight height reportedWeight reportedHeight
195    F      62    164              61             161
196    M      74    175              71             175
197    M      83    180              80             180
198    M      81    175              NA              NA
199    M      90    181              91             178
200    M      79    177              81             178
> mean(Davis$height)
[1] 170.565
```

Fichier STT1000.R

Les fichiers .R contiennent des codes R.

```
> url = "http://freakonometrics.free.fr/STT1000.R"
> download.file(url,"STT1000.R")
> source("STT1000.R")
```



Base du langage R

```
> sum(Davis$sex=="M")
[1] 88
> mean(Davis$height[Davis$sex=="M"])
[1] 178.0114
> mean(Davis$height[Davis$sex=="F"])
[1] 164.7143
```

```
> aggregate(Davis$height, by=list(Davis$sex),FUN=mean)
  Group.1        x
1       F 164.7143
2       M 178.0114
```

```
> tapply(Davis$height, Davis$sex, mean)
       F        M
164.7143 178.0114
```

Pour un cours de R, parcourir le livre d'Ewen Gallic.

Base du langage R

Par exemple, on peut programmer une descente de gradient,

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \text{ à partir d'une valeur initiale } x_0$$

```
> v = c(0.89367,-1.04729,1.97133,-0.38363,1.65414)
> f = function(x) sum((v-x)^2)
> df = function(x) -2*sum((v-x))
> d2f = function(x) 2*length(v)
> x = rep(2,100)
> for(i in 2:100) x[i]=x[i-1]-df(x[i-1])/d2f(x[i-1])
> x[100]
[1] 0.617644
```

ou

```
> df = function(x, h=1e-5) (f(x+h)-f(x))/h
> d2f = function(x, h=1e-5) (df(x+h)-df(x))/h
```

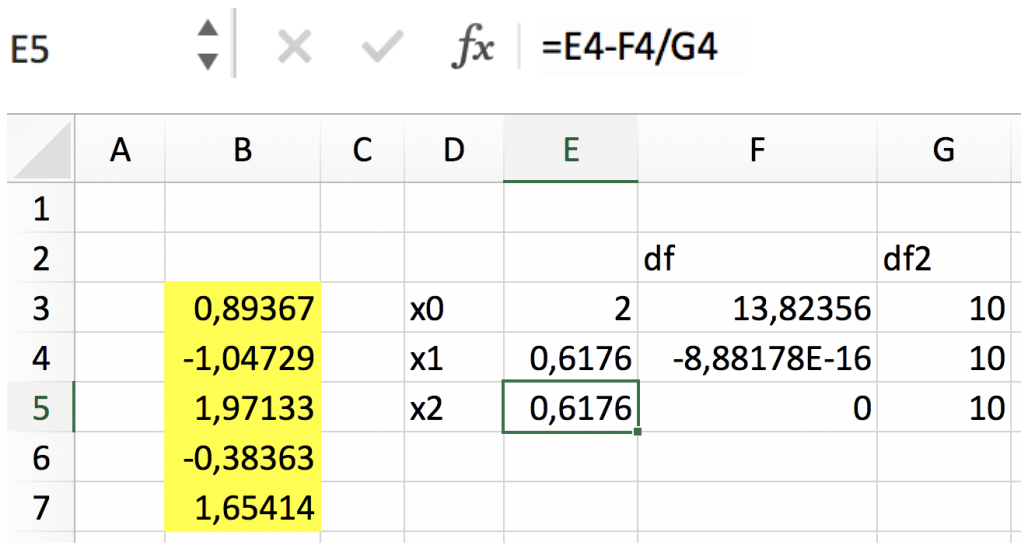
les fonctions, les boucles... un peu hors sujet ici, mais indispensable

Et excel ?

A priori excel peut servir pour des statistiques descriptives, mais ce n'est pas ce qu'on va faire dans ce cours !

Avec un peu de courage, il doit être possible de refaire l'intégralité de ce que l'on verra dans un classeur excel...
Par exemple, on peut "programmer" une descente de gradient,

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \text{ à partir d'une valeur initiale } x_0$$



	A	B	C	D	E	F	G
1							
2						df	df2
3		0,89367		x0	2	13,82356	10
4		-1,04729		x1	0,6176	-8,88178E-16	10
5		1,97133		x2	0,6176	0	10
6		-0,38363					
7		1,65414					

Et excel ?

Mais les formules sont cachées dans les cellules, ce qui rend un classeur excel difficilement auditable

	A	B	C	D	E	F	G
1							
2						df	df2
3		0,89367		x0	2	2,21266	10
4		-1,04729		x1	1,7787	5,652048	10
5		1,97133		x2	1,2135	-1,5156016	10
6		-0,38363		x3	1,3651	3,49743872	10
7		1,65414		x4	1,0153	-1,277589024	10
8				x5	1,1431	#VALEUR!	10

cf https://www.wired.co.uk/article/spreadsheet-excel-errors
ou https://www.bbc.com/news/magazine-22213219

"*teaching Excel in actuarial undergraduate programs is important, but only in courses dedicated to operational risk*"