ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Εισαγωγή στην R, εγκατάσταση, βασική χρήση, μεταβλητές, τελεστές, δεδομένα

Πέτρος Κολοβός

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΓΕΝΕΤΙΚΗΣ

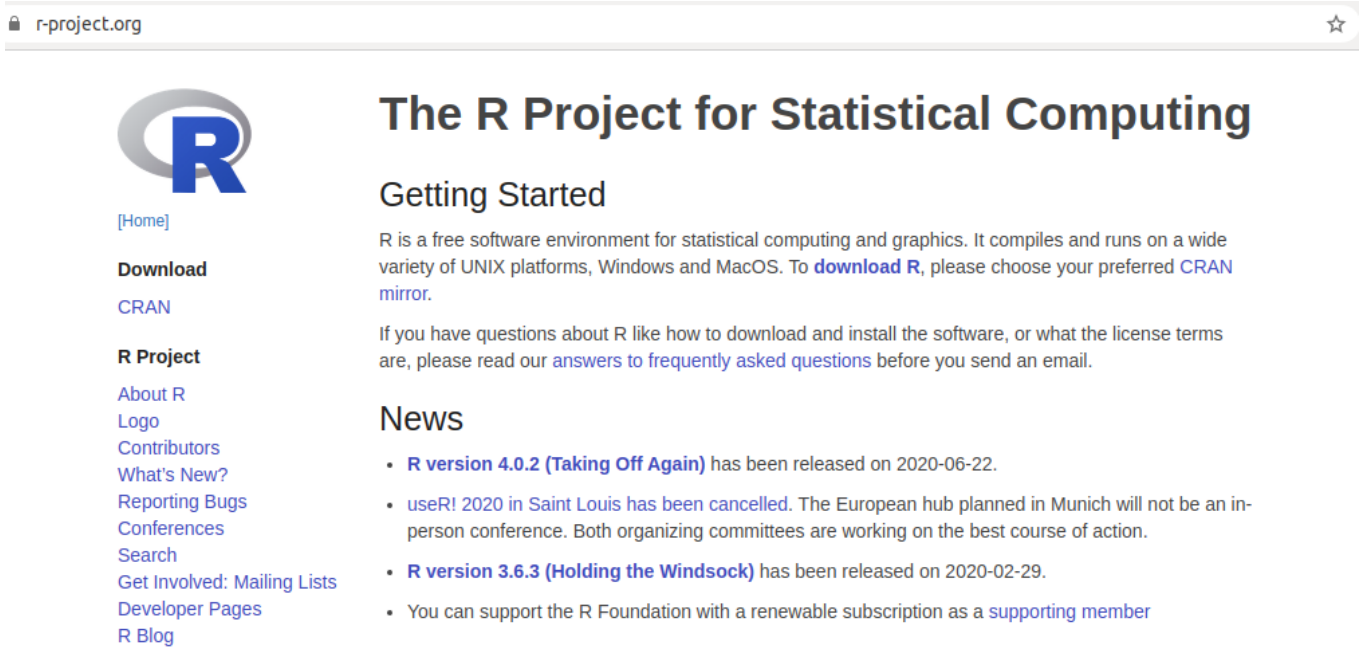
R

- Η R είναι μια γλώσσα προγραμματισμού που βασίστηκε στην S που αναπτύχθηκε στα Bell Laboratories (πρώην AT & T, και πλέον πλέον Lucent Technologies) από τον John Chambers και τους συναδέλφους του.

- Είναι ταυτόχρονα μια γλώσσα κι ένα περιβάλλον προγραμματισμού. Επιτρέπει τόσο τη συγγραφή και εκτέλεση προγραμμάτων αλλά και την απευθείας εκτέλεση εντολών στο περιβάλλον (ή αλλιώς την “κονσόλα”).

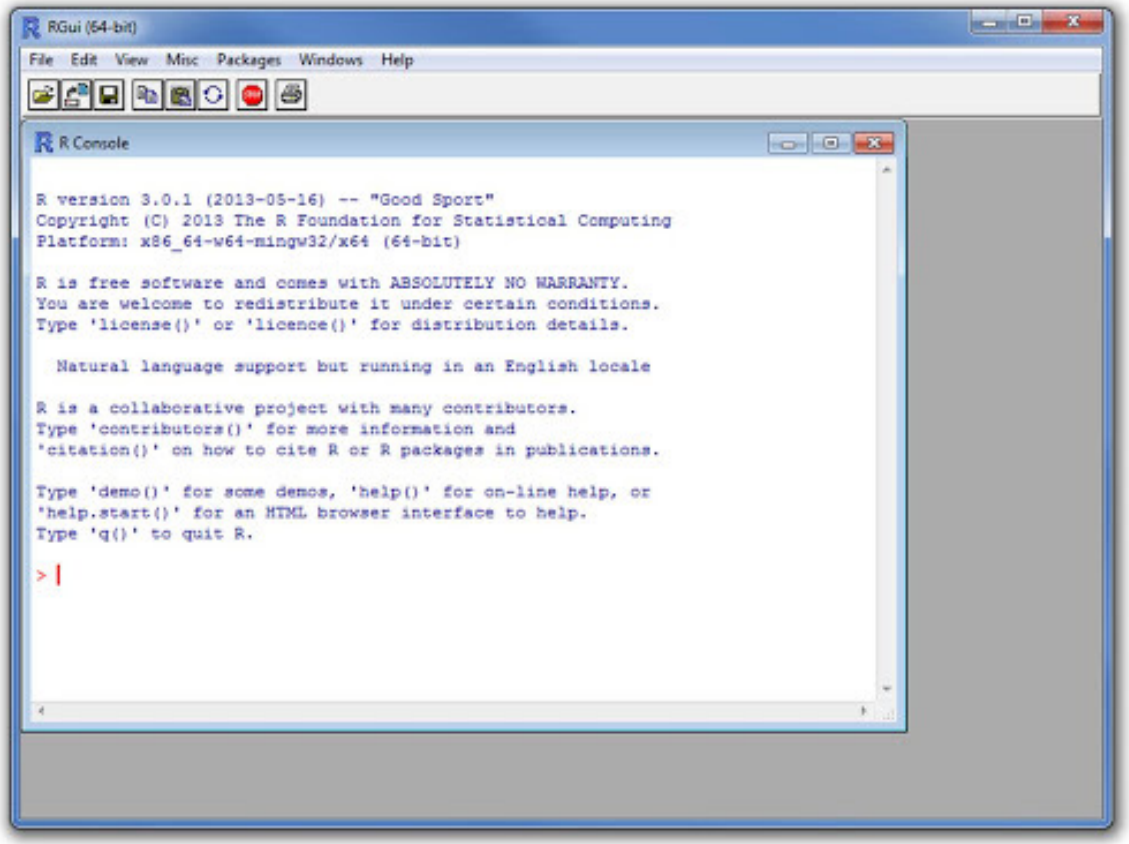
Εγκαθιστώντας την R

- Η R διατίθεται δωρεάν κάτω από μια άδεια γενικής χρήσης (GNU GPL v2). Η εγκατάσταση της R είναι εξαιρετικά απλή και γίνεται μέσα από την ιστοσελίδα του R Project (https://www.r-project.org/)
- Στην συνέχεια και ανάλογα με το λειτουργικό σύστημα ο χρήστης επιλέγει την κατάλληλη έκδοση και προχωράει στην εγκατάσταση.



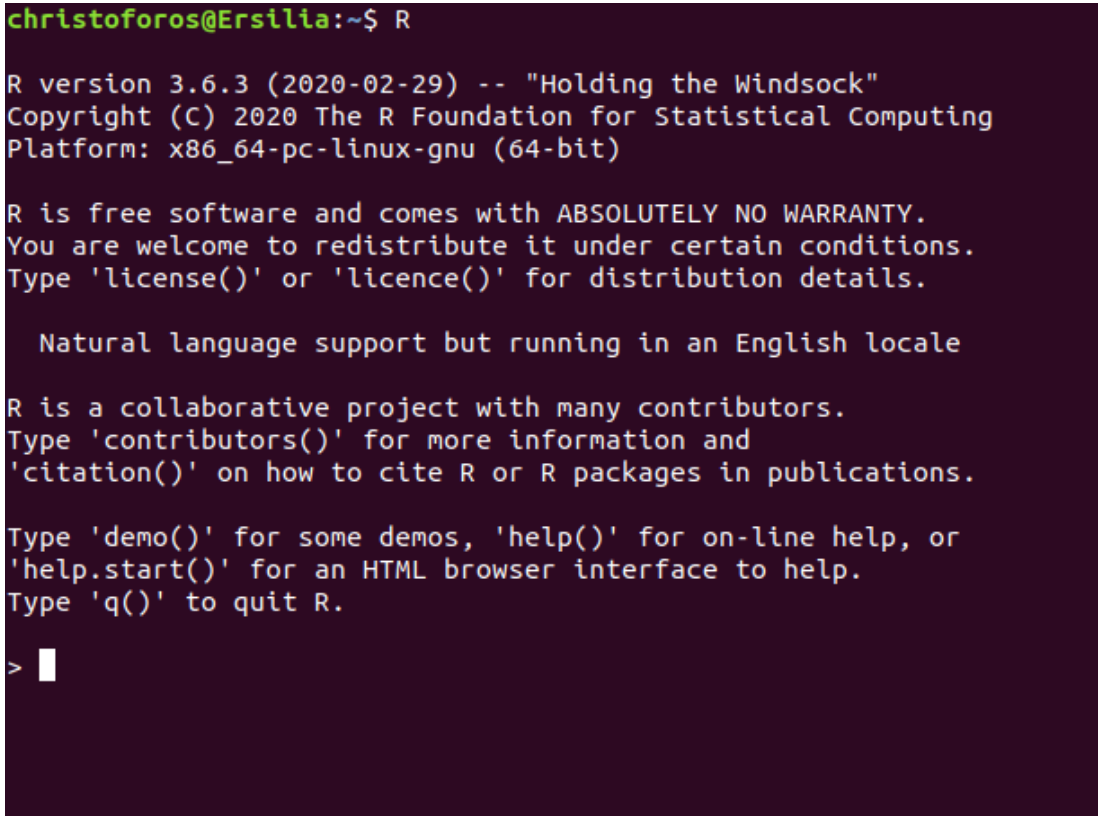
Εκκινώντας την R

- Η εκκίνηση της R σε Windows PC ή σε Mac γίνεται με την απλή εκτέλεση του προγράμματος (διπλό κλικ στο εικονίδιο της R).



Εκκινώντας την R

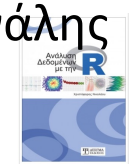
- Σε λειτουργικό σύστημα γραμμής εντολών (π.χ. Linux) η εκκίνηση γίνεται από τη γραμμή εντολών με απλή πληκτρολόγηση του κεφαλαίου R. Εδώ δεν υπάρχει γραφικό interface και όλες οι εντολές δίνονται από τη γραμμή εντολών που όμως συμπεριφέρεται σαν “κονσόλα”.



Πλεονεκτήματα της R

Υπάρχει μια πληθώρα λόγων για τους οποίους κανείς αξίζει να επιλέξει την R για την ανάλυση δεδομένων. Πέρα από το γεγονός ότι **υποστηρίζεται από ένα ευρύτατο δίκτυο συγγραφέων εφαρμογών και χρηστών**, είναι διαθέσιμη **δωρεάν** και μπορεί να **εγκατασταθεί ταχύτατα**, κάποια πιο σημαντικά πλεονεκτήματα είναι τα παρακάτω:

- **Απλότητα**. Η R είναι εξαιρετικά απλή στη χρήση ακόμα και για κάποιον που δεν έχει προηγούμενη εμπειρία στον προγραμματισμό. Τα ολοκληρωμένα περιβάλλοντα όπως το R Studio (βλ. παρακάτω) επιτρέπουν τον εύκολο χειρισμό εντολών ενώ το interface επιτρέπει την παρακολούθηση των αποτελεσμάτων του κώδικα με άμεσο τρόπο κι έτσι την ευκολότερη διόρθωση σφαλμάτων (debugging).
- **Γρήγορη καμπύλη εκμάθησης**. Για τους παραπάνω λόγους η R είναι κατά πάσα πιθανότητα η γλώσσα με την ταχύτερη καμπύλη εκμάθησης. Είναι εξαιρετικά εύκολο για έναν αρχάριο να πραγματοποιήσει αρκετά πολύπλοκες αναλύσεις ακολουθώντας μια σειρά από απλά βήματα όπως ελπίζουμε να διαπιστώσετε και μόνοι σας ολοκληρώνοντας το πρώτο μέρος αυτού του βιβλίου.
- **Ευελιξία και Πληρότητα**. Κανείς δεν φιλοδοξεί να μάθει όλες τις λεπτομέρειες μιας γλώσσας προγραμματισμού πριν προσπαθήσει να την εφαρμόσει για την επίλυση προβλημάτων που τον ενδιαφέρουν. Η R προσφέρει μεγάλη ευελιξία σε αυτό το επίπεδο καθώς οι λειτουργίες της είναι οργανωμένες σε **πακέτα/βιβλιοθήκες** συναρτήσεων που είναι εξειδικευμένα για συγκεκριμένα προβλήματα. Ο χρήστης έτσι μπορεί να καταφύγει σε αυτά χωρίς να χρειαστεί να εξοικειωθεί με το σύνολο των διαθέσιμων εφαρμογών. Από την άλλη πλευρά, ο μεγάλος αριθμός των χρηστών της R σε συνδυασμό με την **κουλτούρα "διαμοιρασμού"** για δωρεάν παροχή εφαρμογών στην κοινότητα (shareware) εξασφαλίζει σε μεγάλο βαθμό τη διαθεσιμότητα μιας πολύ μεγάλης ποικιλίας συναρτήσεων και μεθόδων για ένα μεγάλο αριθμό προβλημάτων.



Εξοικείωση με την R

Μπορεί να τυπώσει μια σειρά χαρακτήρων μέσα σε διπλά ή μονά εισαγωγικά:

Hide

```
"my first R string"
```

```
## [1] "my first R string"
```


Εξοικείωση με την R

Μπορεί κανείς να κάνει μια αριθμητική πράξη στην R χρησιμοποιώντας τα γνωστά σύμβολα πατώντας enter μετά το τέλος της παράσταση

Hide

```
5+10
```

```
## [1] 15
```


Εξοικείωση με την R

Η R βασίζεται σε μια πολύ μεγάλη γκάμα συναρτήσεων που μπορούν να εκτελέσουν από τις πιο απλές, έως τις πιο σύνθετες διαδικασίες.
Η συνάρτηση που δίνει την τετραγωνική ρίζα ενός αριθμού είναι η sqrt() και εφαρμόζεται με τον αριθμό του οποίου τη ρίζα αναζητούμε μέσα σε παρένθεση ως εξής:

Hide

```
sqrt(12)
```

```
## [1] 3.464102
```


Εξοικείωση με την R

Η R μας επιτρέπει να εξετάσουμε το εγχειρίδιο χρήσης κάθε συνάρτησης που είναι καταγεγραμμένο σε ειδικό αρχεία βοήθειας (help file) για κάθε συνάρτηση.

Hide

```
?sqrt
```

MathFun {base} R Documentation

Miscellaneous Mathematical Functions

Description

abs(x) computes the absolute value of x, sqrt(x) computes the (principal) square root of x, $\sqrt{x}$.

The naming follows the standard for computer languages such as C or Fortran.

Usage

```
abs(x)
sqrt(x)
```

Arguments

x	a numeric or complex vector or array.

Details

These are internal generic primitive functions: methods can be defined for them individually or via the Math group generic. For complex arguments (and the default method), z, abs(z) == Mod(z) and sqrt(z) == z^0.5.

Εξοικείωση με την R

Είναι σημαντικό να γνωρίζουμε σε ποιο φάκελο εργαζόμαστε καθώς όπως θα δούμε σε επόμενο Κεφάλαιο, αυτό καθορίζει τον τρόπο με τον οποίο διαβάζουμε και γράφουμε σε αρχεία. Για να βεβαιωθούμε για τον φάκελο εργασίας πληκτρολογούμε:

Hide

```
getwd()
```

Hide

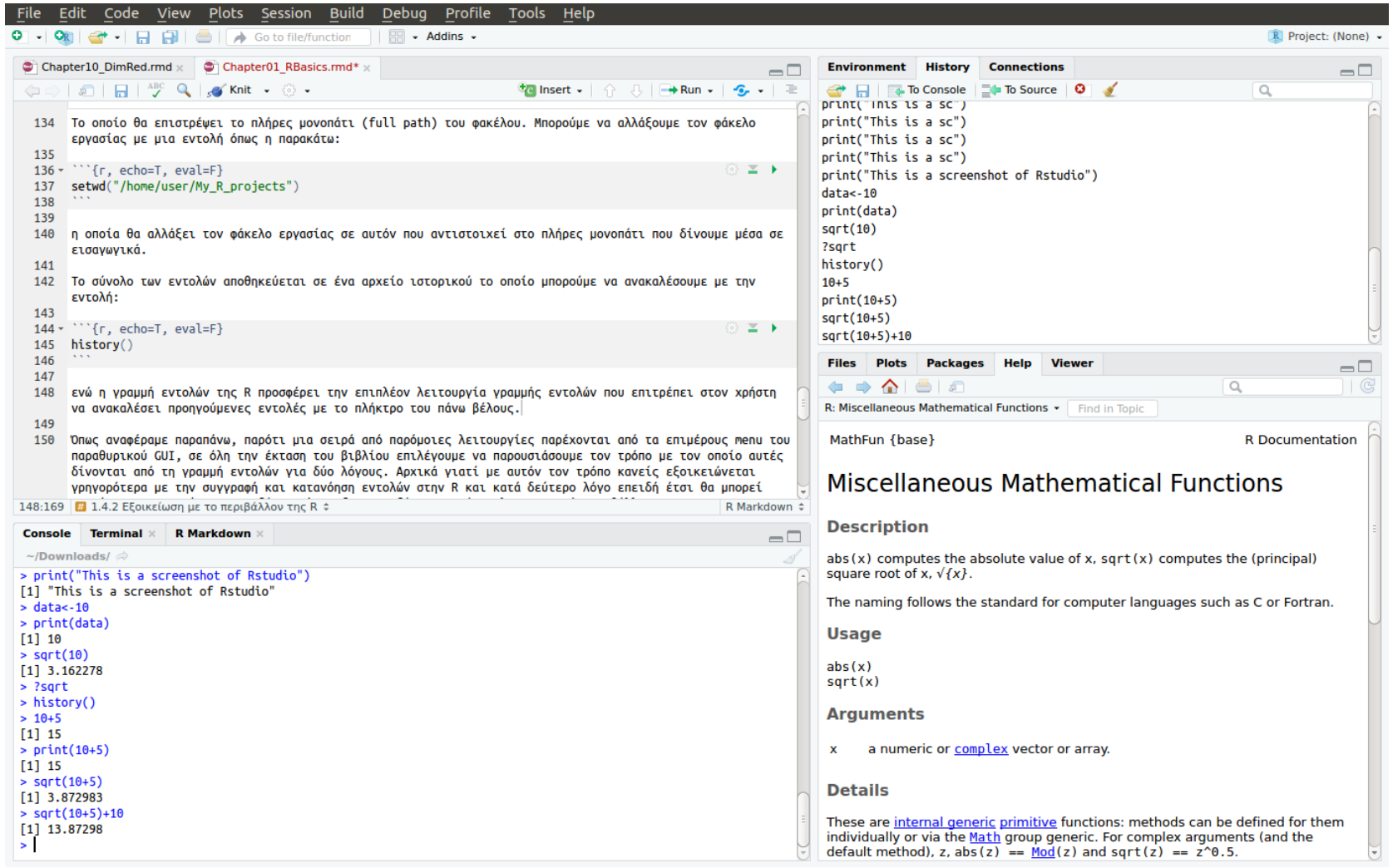
```
setwd("/home/user/My_R_projects")
```

Ολοκληρωμένα Περιβάλλοντα (Integrated Environments)

- Παρότι ο καλύτερος τρόπος να εξοικειωθεί κανείς με την R είναι μέσω της γραμμής εντολών από την κονσόλα, υπάρχουν γραφικά interface που λειτουργούν ως ολοκληρωμένα περιβάλλοντα ανάπτυξης εφαρμογών (Integrated Development Environments, **IDE**) και που είναι εξαιρετικά χρήσιμα τόσο για τους αρχάριους όσο και για τους προχωρημένους χρήστες.
- Το πιο ευρέως χρησιμοποιούμενο είναι το **R Studio** (https://www.rstudio.com/), ένα IDE που επιτρέπει στον χρήστη να συντάσσει κώδικα, να κράταει τον έλεγχο των μεταβλητών και των συναρτήσεων που δημιουργεί, να αποθηκεύει γραφικά και να ανατρέχει στο ιστορικό εντολών με έναν εποπτικό τρόπο.
- Το R Studio διαρθρώνεται σε διαφορετικά πλαίσια (panels) που περιέχουν τις εντολές, σχόλια, την κονσόλα, γραφικά κλπ.

R-studio

Στο παράδειγμα διάθρωσης του R Studio στην εικόνα βλέπετε τις δραστηριότητες διαρθρωμένες σε πλαίσια, με το κείμενο και τον κώδικα πάνω αριστερά, την κονσόλα ακριβώς από κάτω και βοηθητικά πλαίσια με το ιστορικό (πάνω δεξιά) και τη βοήθεια συναρτήσεων (κάτω δεξιά).



R-studio

Η εργασία σε ένα IDE, όπως το R Studio επιτρέπει:

1. Την καλύτερη **οργάνωση των εντολών** μαζί με παράλληλο σχολιασμό.
2. Την εκτέλεση τμημάτων κώδικα σε ξεχωριστά πεδία (code chunks), κάτι που μας επιτρέπει να παρακολουθούμε καλύτερα τη **ροή των αναλύσεων**
3. Την πιο γρήγορη αποσφαλμάτωση (**debugging**) αφού μπορούμε να εντοπίσουμε ευκολότερα τα σημεία που υπάρχει πρόβλημα.
4. Την γενικότερη **εποπτεία αναλύσεων, αποτελεσμάτων, μεταβλητών, διαγραμμάτων** κλπ.

R-studio

Η εργασία στο R Studio επιτρέπει όπως είπαμε την παράθεση σχολίων ανάμεσα στον κώδικα.

Για παράδειγμα, οι παρακάτω εντολές δημιουργούν δυο μεταβλητές x, y που είναι διανύσματα (vector) με μήκος 10 και στη συνέχεια τα εκτυπώνουν στην κονσόλα:

Hide

```
x<-seq(1:10)
y<-x**2
print(x); print(y)
```

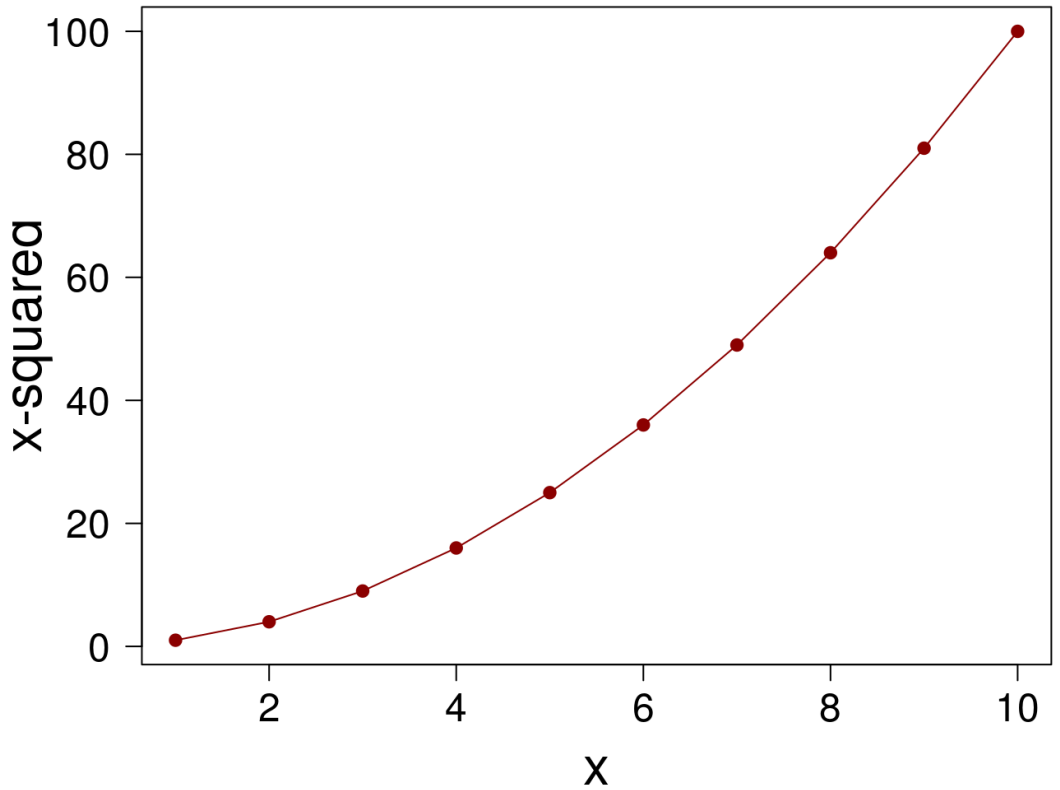
```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
##  [1]   1   4   9  16  25  36  49  64  81 100
```

R-studio

Hide

```
plot(seq(1:10), seq(1:10)**2, type="o", col="dark red", xlab="x", ylab="x-squared", pch=19, cex.
lab=2, cex.axis=1.5, las=1)
```



Ύλη

Ανάλυση δεδομένων με την R

Κεφάλαιο 1

Μεταβλητές στην R

- Όπως σε όλες τις γλώσσες προγραμματισμού, έτσι και στην R οι **μεταβλητές** βρίσκονται στο επίκεντρο κάθε ανάλυσης.
- Στο περιβάλλον της R οι μεταβλητές εισάγονται με το όνομά τους χωρίς να χρειάζεται να εξειδικευτεί καμία επιπλέον παράμετρος. Για παράδειγμα, αν γράψουμε στην κονσόλα την παρακάτω εντολή:

Hide

```
x<-5
```

- Θα δημιουργήσουμε μια μεταβλητή που το όνομα της είναι x και θα έχει την τιμή 5.

Hide

```
x
```

```
## [1] 5
```


Εκχώρηση τιμής σε μεταβλητή

- Η παραπάνω πράξη κατά την οποία δημιουργήσαμε μια μεταβλητή με την τιμή 5 πραγματοποιείται με τη χρήση του τελεστή απόδοση τιμής "**->**".

Hide

```
x<-5
```

- Πρόκειται ουσιαστικά για την ένωση των συμβόλων της παύλας (dash) "-" και του μεγαλύτερου ">" (ή μικρότερου ανάλογα με τη φορά της πράξης).
- Με το συγκεκριμένο συνδυασμό συμβόλων δίνουμε την εντολή στην R να δώσει την τιμή που βρίσκεται από την πλευρά της παύλας στην μεταβλητή που βρίσκεται στην κορυφή του βέλους. Έτσι οι δύο παρακάτω εντολές είναι ισοδύναμες

Hide

```
x<-5
5->x
```

- ενω η παρακάτω είναι λάθος καθώς δεν είναι δυνατόν να αποδοθεί τιμή σε μια σταθερά (όπως είναι το 5).

Hide

```
x->5
```

Εκχώρηση τιμής σε μεταβλητή

- Είναι σημαντικό να θυμόμαστε ότι η R δεν ενημερώνει τον χρήστη όταν πρόκειται να γράψει πάνω σε υπάρχουσες μεταβλητές.
- Έτσι αν π.χ. μια μεταβλητή προϋπάρχει, μια νέα πράξη απόδοσης τιμής θα αντικαταστήσει την παλιά τιμή με την καινούργια χωρίς καμία προειδοποίηση.

Hide

```
x<-5
y<-x # το y είναι ίσο με 5
y<-6 # το y γίνεται ίσο με 6
```

Εκχώρηση τιμής σε Κανόνες ονομασίας μεταβλητών

Ως ονόματα μεταβλητών μπορούν να χρησιμοποιηθούν συνδυασμοί χαρακτήρων και αριθμών κάτω από τις παρακάτω προϋποθέσεις:

- Ο πρώτος χαρακτήρας να είναι πάντοτε γράμμα (πεζό ή κεφαλαίο). Κάτω από προϋποθέσεις μπορεί να είναι και η τελεία “.” αλλά καλό είναι να αποφεύγεται.
- Να μην περιέχουν σύμβολα άλλα από την τελεία “.” και το underscore “_“.
- Να μην ταυτίζονται με ονόματα αντικειμένων της R όπως συναρτήσεις ή ειδικές εντολές.

```
total, my_variable0, my.value, a1_number
```

vs

```
23val, _sum, TRUE
```

Είδη μεταβλητών στην R

Η R δεν περιορίζεται στη χρήση αριθμητικών μεταβλητών. Έτσι τα βασικά είδη (classes) μεταβλητών είναι:

- ✓ **Αριθμοί**
- ✓ **Σειρές Χαρακτήρων**
- ✓ **Λογικές τιμές**

Hide

```
an_integer<-12
a_real<-3.141
a_complex<-5-8i
```

- Η R χειρίζεται με τον ίδιο τρόπο τους δύο πρώτους αποδίδοντάς τους σε ένα είδος (class) που ονομάζεται numeric, ενώ κρατά έναν ξεχωριστό είδος που ονομάζεται complex για τον τελευταίο.

Είδη μεταβλητών στην R

- Μπορούμε να δούμε το είδος μιας μεταβλητής χρησιμοποιώντας μια συνάρτηση που ονομάζεται **class()**. Η εφαρμογή συναρτήσεων στην R γίνεται με την αναγραφή του ονόματός της και με το όρισμα της μέσα σε παρένθεση. Για παράδειγμα, αν θέλουμε να μάθουμε το είδος της μεταβλητής an_integer δεν έχουμε παρά να γράψουμε

Hide

```
class(an_integer)
```

```
## [1] "numeric"
```

- Ομοίως για την a_complex:

Hide

```
class(a_complex)
```

```
## [1] "complex"
```

Είδη μεταβλητών στην R

- Οι σειρές χαρακτήρων (character) είναι οποιαδήποτε τιμή δίνεται μέσα σε διπλά ή μονά εισαγωγικά

Hide

```
my_name<-'Christoforos'
my_passion<-"football"
a_character_variable<-"ytg67_a?"
```

- Ο περιορισμός των εισαγωγικών ισχύει για οποιαδήποτε διαδοχή συμβόλων ακόμα κι αν είναι αριθμητικά. Έτσι η παρακάτω μεταβλητή είναι character

Hide

```
secret_variable<-'568'
```


Είδη μεταβλητών στην R

- Εκτός από αριθμούς και χαρακτήρες, η R διακρίνει κι ένα τρίτο είδος μεταβλητών, τις **λογικές τιμές**. Αυτές δίνονται με τις λέξεις TRUE, FALSE ή με τις συντομογραφίες τους T, F. Οι λογικές τιμές δεν δίνονται μέσα σε διπλά εισαγωγικά αλλιώς εκλαμβάνονται από την R ως σειρές χαρακτήρων:

Hide

```
my_logical1<-T
my_logical2<-FALSE
my_not_so_logical<-"FALSE"
```

Hide

```
class(my_not_so_logical)
```

```
## [1] "character"
```

vs

Hide

```
class(my_logical1)
```

```
## [1] "logical"
```

Αριθμητικές Πράξεις

- Οι αριθμητικές πράξεις πραγματοποιούνται με τη χρήση ειδικών συμβόλων που είναι πολύ παρόμοια με αυτά που χρησιμοποιούνται από άλλες γλώσσες. Οι τελεστές για τις βασικές αριθμητικές πράξεις φαίνονται στον παρακάτω πίνακα

Σύμβολο	Πράξη
+	Πρόσθεση
-	Αφαίρεση
*	Πολλαπλασιασμός
/	Διαίρεση
**	Ύψωση σε δύναμη
%%	Υπόλοιπο διαίρεσης

Hide

```
6+7
```

```
## [1] 13
```

Αριθμητικές Πράξεις

Hide

```
d<-3
h<-2
surface<-(2*3.14*(d/2)**2)+(2*3.14*d*h)
```

Λογικές πράξεις

- Οι λογικές πράξεις αφορούν τη σύγκριση ομοιότητας/διαφοράς και (για αριθμητικές μόνο μεταβλητές) τις συγκρίσεις μεγαλύτερης/μικρότερης τιμής. Οι τελεστές λογικών πράξεων είναι:

Σύμβολο	Πράξη
==	Ίσο/Όμοιο
!=	Διάφορο
>	Μεγαλύτερο
<	Μικρότερο
>=	Μεγαλύτερο ή ίσο
<=	Μικρότερο ή ίσο

Hide

```
x<-6
y<-8
x>y
```

```
## [1] FALSE
```

Συναρτήσεις

- Ένα από τα χαρακτηριστικά της R που την κάνουν τόσο εύχρηστη είναι η **ύπαρξη έτοιμων συναρτήσεων** που πραγματοποιούν πολύπλοκες υπολογιστικές διεργασίες χωρίς να χρειάζεται να κωδικοποιηθούν από τον χρήστη.
- Οι συναρτήσεις είναι αντικείμενα της R που δρουν πάνω σε **μία ή περισσότερες μεταβλητές**. Η εφαρμογή τους γίνεται με τη αναγραφή του ονόματος της συνάρτησης και την τοποθέτηση της μεταβλητής-ορίσματος μέσα σε παρενθέσεις.
- Πολύ συχνά οι συναρτήσεις **δέχονται επιπλέον παραμέτρους** που χωρίζονται με κόμμα μέσα στο πλαίσιο των παρενθέσεων και πάντα μετά τις μεταβλητές.
- Η γενική σύνταξη είναι η εξής

```
fun(variable(s), parameters)
```


Δημιουργία συναρτήσεων

- R είναι μια γλώσσα προγραμματισμού κι ως τέτοια επιτρέπει στον χρήστη να δημιουργήσει δικές του συναρτήσεις συνδυάζοντας ήδη υπάρχουσες με στοιχεία προγραμματισμού.
- Ας υποθέσουμε για παράδειγμα ότι θέλουμε να γράψουμε μια συνάρτηση για την εκτέλεση της πράξης της πρόσθεσης δύο αριθμών. Η σύνταξη για την δημιουργία της συνάρτησης είναι η εξής.

Hide

```
sumAB<-function(a,b){
  result<-a+b
  return(result)
  }
```

Hide

```
sumAB(12,28)
```

```
## [1] 40
```

Έτοιμα (built-in) δεδομένα

- Για πρακτικούς λόγους η R περιέχει **προεγκατεστημένα σύνολα δεδομένων**. Κάποια από αυτά είναι μέρος της βασικής έκδοσης της R και κάποια άλλα ανήκουν σε ειδικές βιβλιοθήκες συναρτήσεων.
- Η ύπαρξή τους εξυπηρετεί εκπαιδευτικούς σκοπούς καθώς χρησιμοποιούνται για την επίδειξη λειτουργιών, συναρτήσεων κλπ.

data()

```
Data sets in package 'datasets':

AirPassengers                Monthly Airline Passenger Numbers 1949-1960
BJsales                      Sales Data with Leading Indicator
BJsales.lead (BJsales)       Sales Data with Leading Indicator
BOD                          Biochemical Oxygen Demand
CO2                          Carbon Dioxide Uptake in Grass Plants
ChickWeight                  Weight versus age of chicks on different diets
DNase                        Elisa assay of DNase
EuStockMarkets               Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde                 Determination of Formaldehyde
HairEyeColor                 Hair and Eye Color of Statistics Students
Harman23.cor                 Harman Example 2.3
Harman74.cor                 Harman Example 7.4
Indometh                     Pharmacokinetics of Indomethacin
InsectSprays                 Effectiveness of Insect Sprays
JohnsonJohnson               Quarterly Earnings per Johnson & Johnson Share
LakeHuron                    Level of Lake Huron 1875-1972
LifeCycleSavings             Intercountry Life-Cycle Savings Data
Loblolly                     Growth of Loblolly pine trees
Nile                         Flow of the River Nile
Orange                       Growth of Orange Trees
OrchardSprays                Potency of Orchard Sprays
PlantGrowth                  Results from an Experiment on Plant Growth
Puromycin                    Reaction Velocity of an Enzymatic Reaction
Seatbelts                    Road Casualties in Great Britain 1969-84
Theoph                       Pharmacokinetics of Theophylline
Titanic                      Survival of passengers on the Titanic
ToothGrowth                  The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions                Student Admissions at UC Berkeley
UKDriverDeaths               Road Casualties in Great Britain 1969-84
UKgas                        UK Quarterly Gas Consumption
USAccDeaths                  Accidental Deaths in the US 1973-1978
USArrests                    Violent Crime Rates by US State
USJudgeRatings               Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure        Personal Expenditure Data
UScitiesD                    Distances Between European Cities and Between US Cities
VADeaths                     Death Rates in Virginia (1940)
WWWusage                     Internet Usage per Minute
WorldPhones                  The World's Telephones
ability.cov                  Ability and Intelligence Tests
airmiles                     Passenger Miles on Commercial US Airlines, 1937-1960
airquality                   New York Air Quality Measurements
anscombe                     Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu                       The Joyner-Boore Attenuation Data
attitude                     The Chatterjee-Price Attitude Data
austres                      Quarterly Time Series of the Number of Australian Residents
beaver1 (beavers)            Body Temperature Series of Two Beavers
beaver2 (beavers)            Body Temperature Series of Two Beavers
```

Έτοιμα (built-in) δεδομένα

Hide

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa      30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora     15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E        21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Έτοιμα (built-in) δεδομένα

- Ένας τρόπος για να καταλάβουμε τη δομή λίγο καλύτερα είναι εφαρμόσουμε μια συνάρτηση που θα μας δώσει τις διαστάσεις της μεταβλητής
- Η dim() δίνει τις διαστάσεις μιας μεταβλητής που στην προκειμένη περίπτωση είναι 32X11, δηλαδή πρόκειται για έναν πίνακα 32 γραμμών και 11 στηλών.

Hide

```
dim(mtcars)
```

```
## [1] 32 11
```

Έτοιμα (built-in) δεδομένα

- Στην περίπτωση που θέλουμε να πάρουμε μια εικόνα για μια σύνθετη μεταβλητή χωρίς να την τυπώσουμε ολόκληρη μπορούμε να χρησιμοποιήσουμε συναρτήσεις που μας δίνουν ένα τμήμα των δεδομένων
- Η **head()** επιστρέφει τις πρώτες γραμμές της μεταβλητής (με προεπιλεγμένη τιμή τις 6 γραμμές). Αντίστοιχα η **tail()** επιστρέφει τις τελευταίες γραμμές. Αν κανείς θέλει να δει έναν συγκεκριμένο αριθμό γραμμών θα πρέπει να προσθέσει μια παράμετρο ακόμα στην εκτέλεση της συνάρτησης:

Hide

```
head(mtcars, n=10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

Έτοιμα (built-in) δεδομένα

- Μια πιο αναλυτική εικόνα για μια οποιαδήποτε μεταβλητή μπορούμε να πάρουμε με τη συνάρτηση **str()**
- Η str() (που είναι συντομογραφία του structure) δίνει τη δομή της μεταβλητής με πολύ πιο αναλυτικό τρόπο από την dim().
- Συγκεκριμένα δίνει αρχικά τον τύπο της μεταβλητης, ο οποίος είναι data.frame δηλαδή πίνακας και οι διαστάσεις του είναι 32X11 (32 παρατηρήσεις για 11 μεταβλητές). Οι 11 "μεταβλητές" εδώ είναι οι παράμετροι για τις οποίες έχουμε μετρήσεις και οι οποίες δίνονται με τα ονόματά τους (mpg, cyl κλπ), καθώς και το είδος της κάθεμίας (εδώ όλες είναι numeric).

Hide

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```


Εισαγωγή Δεδομένων από αρχεία

- Τα προεγκατεστημένα σύνολα δεδομένων είναι χρήσιμα για τη δοκιμή λειτουργιών και συναρτήσεων, ωστόσο ο σκοπός μας είναι να αναλύσουμε δεδομένα που έχουν προκύψει από δικά μας πειράματα και μετρήσεις. Προκειμένου να συμβεί αυτό θα πρέπει τα εξωτερικά αντικείμενα/αρχεία που περιέχουν τα αποτελέσματά μας να περάσουν στην R και να αποθηκευτούν σε μεταβλητές. Η R υποστηρίζει αυτή τη διαδικασία με μια σειρά από **συναρτήσεις ανάγνωσης αρχείων.**

Συνάρτηση	Λειτουργία
read.table()	Ανάγνωση Πίνακα
read.csv()	Ανάγνωση Πίνακα με διαχωριστή κόμμα
read.delim()	Ανάγνωση Πίνακα με δήλωση διαχωριστή
readLines()	Ανάγνωση κειμένου ανα γραμμή

Εισαγωγή Δεδομένων από αρχεία

- Η γενική σύνταξη των παρακάτω συναρτήσεων είναι η εξής

```
readfunction("nameoffile", header=T/F, sep="character")
```

✓ Στο **nameoffile** δίνεται το όνομα του αρχείου (μαζί με την κατάληξη) μέσα σε εισαγωγικά. Ανάλογα με τον φάκελο directory στο οποίο δουλεύουμε ενδέχεται να πρέπει να δώσουμε όχι μόνο το όνομα του αρχείου αλλά και το πλήρες μονοπάτι (full path) δηλαδή όλη τη διεύθυνση του αρχείου στον υπολογιστή που εργαζόμαστε. Εναλλακτικά μπορούμε να αλλάξουμε τον φάκελο εργασίας με τη συνάρτηση setwd().

✓ Στο **header=LOGICAL** δίνεται μια λογική τιμή TRUE/FALSE ή πιο σύντομα T/F η οποία καθορίζει αν η πρώτη γραμμή του αρχείου θα χρησιμοποιηθεί ως επικεφαλίδα ή όχι.

✓ Στο **sep="character"** δίνεται ως τιμή ένας χαρακτήρας (ή μια σειρά χαρακτήρων) που θα χρησιμοποιηθεί ως διαχωριστής. Αν π.χ. το αρχείο που θέλουμε να διαβάσουμε είναι tsv (tab separated values) τότε η παράμετρος θα είναι **sep="\t"**.

Ανάγνωση Δεδομένων σε Πίνακα

- Ας δούμε τη λειτουργία μιας read συνάρτησης με ένα παράδειγμα. Στον φάκελο στον οποίο εργαζόμαστε υπάρχει ένα αρχείο που ονομάζεται Arthritis.tsv και είναι, όπως φαίνεται από την κατάληξη του, tab separated. Για να το διαβάσουμε με την read.delim() μπορούμε να γράψουμε

Hide

```
data<-read.delim("Arthritis.tsv", header=T, sep="\t")
```

Hide

```
str(data)
```

```
## 'data.frame':    84 obs. of  5 variables:
##  $ ID       : int  57 46 77 17 36 23 75 39 33 55 ...
##  $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Age      : int  27 29 30 32 46 58 59 NA 63 63 ...
##  $ Improved : Factor w/ 3 levels "Marked","None",..: 3 2 2 1 1 1 2 1 2 2 ...
```


Ορισμός φακέλου εργασίας και πλήρους μονοπατιού

- Στα παραδείγματα που είδαμε παραπάνω τα αρχεία που διαβάσαμε βρίσκονται στο φάκελο εργασίας, δηλαδή τον φάκελο στον οποίο έχουμε ανοίξει την R. Όπως είδαμε στο προηγούμενο κεφάλαιο, για να δούμε ποιος είναι ο φάκελος εργασίας εκτελούμε την

Hide

```
getwd()
```

```
## [1] "/home/christoforos/Dropbox/DataAnalysisRBook"
```

- Να αλλάξουμε το φάκελο εργασίας, με τη συνάρτηση setwd()

Hide

```
setwd("~/Documents")
getwd()
```

```
## [1] "/home/christoforos/Documents"
```

- Να δώσουμε το πλήρες μονοπάτι του αρχείου ακολουθώντας την ιεραρχία των φακέλων

Hide

```
data<-read.delim("~/Dropbox/DataAnalysisRBook/Arthritis.tsv", header=T, sep="\t")
```

Εγγραφή Αποτελεσμάτων σε αρχεία

- Η R επιτρέπει εκτός από την εισαγωγή και την εξαγωγή δεδομένων σε αρχεία στον υπολογιστή, με μια σειρά από συναρτήσεις τύπου **write().**

Hide

```
setwd("~/Dropbox/DataAnalysisRBook/")
write(data, file="out.txt")
```

- Η οποία μπορεί να γίνει πολύ πιο περίπλοκη εάν προστεθούν επιπλέον επιλογές στην εντολή

Hide

```
write(data, file="out.txt", append=T, sep="\t", ncolumns=5)
```

- Εκτέλεση της **write()** με την παραπάνω σύνταξη δεν θα γράψει μόνο τα δεδομένα σε ένα όνομα αρχείου "out.txt", αλλά θα προσθέσει τα δεδομένα στο τέλος αυτού του αρχείου εάν υπάρχει ήδη (**append=T**). Επιπλέον, τα δεδομένα θα γραφτούν σε 5 στήλες χωρισμένες με tab.

Εγγραφή Αποτελεσμάτων σε αρχεία

- Σε ό,τι αφορά δεδομένα η καλύτερη επιλογή είναι η **write.table()**, η οποία επιτρέπει στον χρήστη να έχει τον καλύτερο έλεγχο της καταγραφής των δεδομένων. Μια ενδεικτική σύνταξή της είναι

Hide

```
write.table(data, file="out.txt", sep="\t", row.names=F, col.names=T, quote=F)
```

- Η παράμετρος **row.names=F** λέει στην R να παραλείψει να απαριθμήσει τις σειρές, καθώς συνήθως προσθέτει και επιπλέον στήλη στο αρχείο εξόδου.
- Η **quote=F** ορίζει ότι οι σειρές χαρακτήρων θα τυπωθούν “γυμνές” στο αρχείο δηλαδή χωρίς να εσωκλείονται σε εισαγωγικά
- Η **sep=“\t”** ορίζει ότι τα δεδομένα θα γραφτούν σε στήλες χωρισμένες με tab

Ανάγνωση έτοιμου κώδικα R

- Η R επιτρέπει την εισαγωγή και εκτέλεση εντολών στην R από αρχεία που είναι αποθηκευμένα στον υπολογιστή μας με τη χρήση της συνάρτησης **source()**. Το όρισμα της source() είναι ένα αρχείο που περιέχει αποκλειστικά εντολές στην R οι οποίες εισάγονται στην κονσόλα ως εξής

Hide

```
source("Rcommands.R")
```


Ύλη

Ανάλυση δεδομένων με την R

Κεφάλαιο 2