Οδηγίες χρήσης του R, μέρος 2^{ο}

Ελληνικά

Αν προσπαθήσουμε να γράψουμε ελληνικά ή να ανοίξουμε κάποιο αρχείο δεδομένων με ελληνικούς χαρακτήρες στο R, μπορεί αντί για ελληνικά να δούμε λατινικούς χαρακτήρες με τόνους ή άλλα καλλικαντζαράκια. Τότε δίνουμε την παρακάτω εντολή για να γυρίσει το R στα ελληνικά:

```
> Sys.setlocale("LC_CTYPE","Greek")
[1] "Greek_Greece.1253"
```

Η απόκριση του R (με μπλε) επιβεβαιώνει τη ρύθμιση των ελληνικών.

Πλαίσια δεδομένων

Το R διατηρεί μετρήσεις μέσα σε δομές που ονομάζονται «πλαίσια δεδομένων» (data frame). Κάθε πλαίσιο δεδομένων περιέχει μία ή περισσότερες μεταβλητές.

Για παράδειγμα, ας καταγράψουμε το φύλο και την ηλικία δύο ατόμων, του Γιάννη και της Μαρίας, σε ένα πλαίσιο δεδομένων το οποίο αναθέτουμε σε μια μεταβλητή με όνομα `atoma`:

```
> atoma<-data.frame(sex=c("M","F"),age=c(21,22))
```

Με τη συνάρτηση `data.frame` ορίζουμε ένα πλαίσιο δεδομένων. Στη συνάρτηση δίνουμε ως ορίσματα τις μεταβλητές που θέλουμε να περιέχει το πλαίσιο, δηλαδή `sex` (φύλο) και `age` (ηλικία). Σε κάθε μεταβλητή δίνουμε τις αντίστοιχες μετρήσεις, ως ακολουθία τιμών (χρησιμοποιώντας τη συνάρτηση `c` που είδαμε στο πρώτο μέρος των οδηγιών).
Αν θέλουμε (δεν είναι υποχρεωτικό) μπορούμε να προσθέσουμε ετικέτες στις σειρές του πλαισίου για να αναγνωρίζουμε ονομαστικά τα δεδομένα:

```
> rownames(atoma)<-c("Γιάννης","Μαρία")
```

Τώρα μπορούμε να δούμε τα περιεχόμενα του πλαισίου δεδομένων atoma:

```
> atoma
        sex age
Γιάννης   M  21
Μαρία     F  22
```

Το R μας δίνει, σε μορφή πίνακα, όλα τα δεδομένα του πλαισίου. Για να εξετάσουμε τη δομή του πλαισίου μπορούμε να χρησιμοποιήσουμε τη συνάρτηση `str`, η οποία μας δίνει περιληπτικά το είδος των δεδομένων και ενδεικτικές τιμές για κάθε στήλη (μεταβλητή):

```
> str(atoma)
'data.frame':   2 obs. of  2 variables:
 $ sex: Factor w/ 2 levels "F","M": 2 1
 $ age: num  21 22
```

Στην πρώτη σειρά της απόκρισης, η συνάρτηση `str` μας ενημερώνει ότι το `atoma` είναι πλαίσιο δεδομένων (`'data frame'`), το οποίο περιέχει δύο «παρατηρήσεις» (`obs.`=observations) και δύο «μεταβλητές». Λέγοντας «παρατηρήσεις» αναφερόμαστε στις σειρές του πλαισίου, ενώ λέγοντας «μεταβλητές» αναφερόμαστε στις στήλες.
Στις επόμενες δύο σειρές της απόκρισης δίνονται οι πληροφορίες που αφορούν σε καθεμιά μεταβλητή ξεχωριστά:

Η πρώτη μεταβλητή ονομάζεται `sex` και είναι τύπου `Factor`, δηλαδή «παράγοντας». Αυτό, στην ορολογία του R, σημαίνει ότι πρόκειται για *κατηγορική* μεταβλητή. Περιλαμβάνει 2 «επίπεδα» (`levels`), δηλαδή δύο κατηγορίες, οι οποίες ονομάζονται `"F"` και `"M"` (γυναίκες και άντρες). Το `"F"` αναφέρεται πρώτο διότι το R χρησιμοποιεί από μόνο του αλφαβητική σειρά για την αναφορά σε κατηγορίες. Η σειρά περιγραφής της μεταβλητής `sex` ολοκληρώνεται με τα πρώτα στοιχεία της στήλης, δηλαδή τον αριθμό 2 (αναφέρεται στη δεύτερη κατηγορία, `"M"`) και τον αριθμό 1 (πρώτη κατηγορία, `"F"`). Αυτό μας λέει ότι η πρώτη σειρά δεδομένων είναι τύπου `"F"` και η δεύτερη τύπου `"M"`.

Η δεύτερη μεταβλητή ονομάζεται `age` και είναι τύπου `num`, δηλαδή «αριθμητική» (`numeric`). Αυτό, στην ορολογία του R σημαίνει ότι πρόκειται για *ποσοτική* μεταβλητή. Δεν χρειάζονται άλλες διευκρινίσεις, καθώς στις ποσοτικές μεταβλητές τα νούμερα είναι αυτονόητα. Η σειρά ολοκληρώνεται με τα πρώτα στοιχεία της στήλης, δηλαδή τους αριθμούς 21 και 22, οι οποίοι αντιστοιχούν στην πρώτη και τη δεύτερη σειρά δεδομένων, αντίστοιχα.

> Να θυμάστε ότι τα κατηγορικά δεδομένα στο R είναι τύπου **factor** ενώ τα ποσοτικά δεδομένα είναι τύπου **numeric**.

Όπως βλέπουμε στην απόκριση της συνάρτησης `str`, πριν από κάθε μεταβλητή εμφανίζεται ένα δολλάριο (`$`). Το σύμβολο του δολλαρίου στο R χρησιμοποιείται για να δηλώνουμε συγκεκριμένες μεταβλητές μέσα σε πλαίσια δεδομένων. Έτσι, για να αναφερθούμε στις ηλικίες (μεταβλητή `age`) που βρίσκονται μέσα στο πλαίσιο `atoma` γράφουμε

```
> atoma$age
[1] 21 22
```

ενώ για να αναφερθούμε στο φύλο (μεταβλητή `sex`) γράφουμε

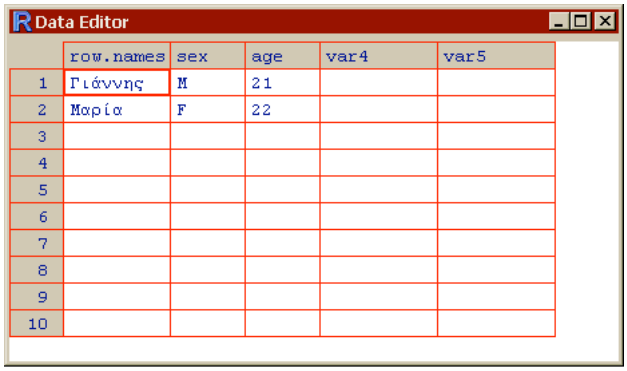
```
> atoma$sex
[1] M F
Levels: F M
```

Στην περίπτωση της κατηγορικής μεταβλητής το R μας ενημερώνει και για το σύνολο των κατηγοριών που περιλαμβάνει η συγκεκριμένη μεταβλητή.
Απομονώνοντας τις μεταβλητές με αυτόν τον τρόπο, μπορούμε να τις χειριστούμε ως κοινές ακολουθίες. Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε τις συναρτήσεις από το πρώτο μέρος των οδηγιών για να υπολογίσουμε το άθροισμα, το πλήθος κλπ.

```
> sum(atoma$age)
[1] 43
> length(atoma$sex)
[1] 2
```

Για την επεξεργασία δεδομένων σε πλαίσια, το R μας δίνει τη συνάρτηση `fix`, με την οποία μας εμφανίζει ένα ειδικό παράθυρο στο οποίο μπορούμε να τροποποιήσουμε ή να προσθέσουμε στοιχεία σε ένα πλαίσιο δεδομένων.

```
> fix(atoma)
```



Όταν τελειώσουμε την προσθήκη ή επεξεργασία των στοιχείων, κλείνουμε το ειδικό παράθυρο με κλικ στο ✕ (πάνω δεξιά γωνία) και η μεταβλητή `atoma` ενημερώνεται αυτόματα.

Γραφική παρουσίαση δεδομένων

Το R διαθέτει πολλές συναρτήσεις για την παρουσίαση και λεπτομερειακή εξέταση και ανάλυση των δεδομένων μας. Ας υποθέσουμε ότι έχουμε πέντε μετρήσεις ύψους ενός ατόμου:

```
> alexh <- c( 1.85, 1.85, 1.81, 1.82, 1.83 )
```

Η μεταβλητή `alexh` περιέχει μια ακολουθία πέντε αριθμών. Με τη συνάρτηση `table` («πίνακας») μπορούμε να μετρήσουμε πόσες φορές εμφανίζεται κάθε τιμή:

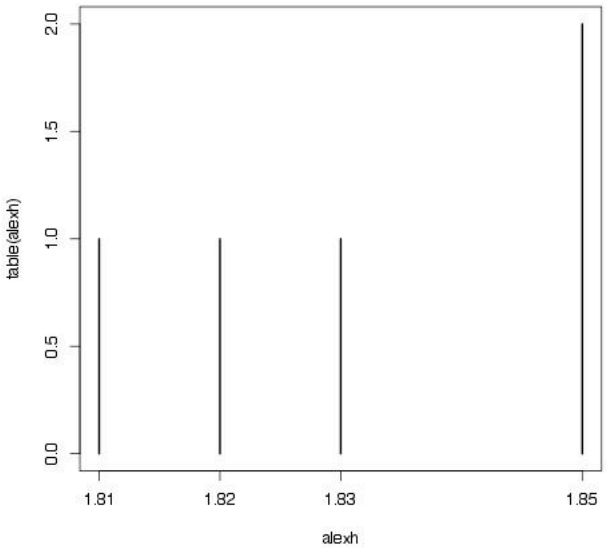
```
> table(alexh)
alexh
1.81 1.82 1.83 1.85
   1    1    1    2
```

Αυτός είναι ένας απλός πίνακας συχνοτήτων. Βλέπουμε ότι η τιμή 1.85 εμφανίζεται δύο φορές ενώ οι άλλες τιμές από μία φορά. Η τιμή που εμφανίζεται τις περισσότερες φορές ονομάζεται «δεσπόζουσα» (mode). Αν η καλύτερη τιμή έβγαινει με ψηφοφορία, η δεσπόζουσα είναι εκείνη που θα κέρδιζε λόγω πλειοψηφίας.

Την πληροφορία αυτή μπορούμε να τη δούμε και γραφικά, με τη συνάρτηση plot:

```
> plot(table(alexh))
```

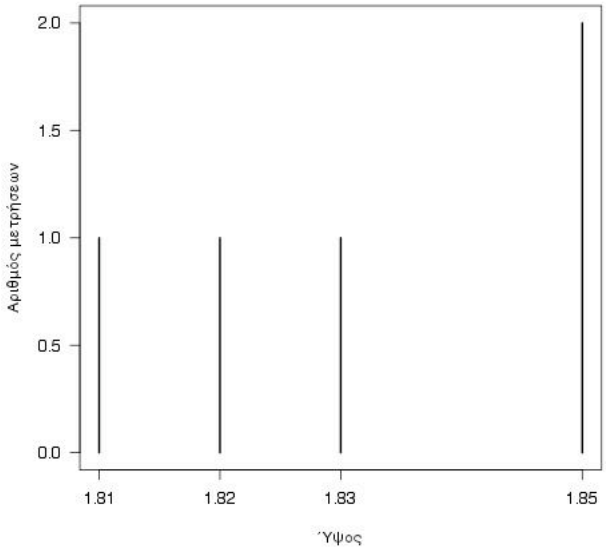
Το R ανοίγει ένα νέο παράθυρο για τη γραφική απεικόνιση, στο οποίο εμφανίζει το εξής :



Εδώ βλέπουμε ένα ραβδόγραμμα με το ύψος στον οριζόντιο άξονα και το πλήθος των αντίστοιχων μετρήσεων στον κατακόρυφο. Κάθε μέτρηση εμφανίζεται σα μια γραμμούλα που φτάνει σε ύψος 1.0, ενώ στο ύψος 1.85, που υπάρχουν δύο μετρήσεις, εμφανίζονται δύο γραμμούλες η μία πάνω στην άλλη, κάνοντας μαζί μια μακρύτερη που φτάνει στο ύψος 2.0. Αυτό είναι ένα διάγραμμα συχνοτήτων, που μας λέει πόσο συχνά εμφανίζεται κάθε αριθμός. Μπορούμε να καλλωπίσουμε κάπως τη γραφική απεικόνιση, προσθέτοντας ετικέτες:

```
> plot(table(alexh),las=1,xlab="Ύψος",ylab="Αριθμός μετρήσεων")
```

Η παράμετρος las στρίβει την αρίθμηση στον κατακόρυφο άξονα ώστε να διαβάζεται όρθια, ενώ οι δύο παράμετροι lab (από το label=ετικέτα) καθορίζουν τις ετικέτες στον οριζόντιο άξονα (με το x) και στον κατακόρυφο άξονα (με το y).



Για να δούμε γραφικά την κατανομή των μετρήσεων στην κλίμακα το R μας δίνει τη συνάρτηση `hist` (histogram=ιστόγραμμα). Την κατανομή αυτή, με περισσότερη αριθμητική λεπτομέρεια αλλά χωρίς γραφικά, μπορούμε να δούμε με τη συνάρτηση `stem` που παράγει διάγραμμα μίσχου-φύλλων.

Για το σχετικό πλήθος των επιμέρους κατηγοριών σε κατηγορικά δεδομένα, έχουμε τη συνάρτηση `table`, που είδαμε παραπάνω ότι μας δίνει τον πίνακα κατανομής, καθώς και τη συνάρτηση `pie`, που μας δίνει γραφικά την ίδια πληροφορία με κυκλικό διάγραμμα (pie chart).

Δοκιμάστε τις συναρτήσεις αυτές στα δικά σας δεδομένα!

Περίληψη δεδομένων

Μια πολύ χρήσιμη συνάρτηση για γρήγορη επισκόπηση των δεδομένων μας είναι η περίληψη:

```
> summary(alexh)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.810   1.820   1.830   1.832   1.850   1.850
```

Τα αποτελέσματα της περίληψης περιλαμβάνουν την ελάχιστη (`Min.`) και μέγιστη (`Max.`) τιμή, το μέσο όρο (`Mean`), καθώς και τρεις ακόμα δείκτες. Ο πιο σημαντικός είναι η διάμεσος (`Median`), δηλαδή η τιμή που είναι μεγαλύτερη από τις μισές μετρήσεις και μικρότερη από τις άλλες μισές. Για να το καταλάβουμε καλύτερα, ας δούμε τις τιμές μας σε αύξουσα σειρά:

```
> sort(alexh)
[1] 1.81 1.82 1.83 1.85 1.85
```

Η μικρότερη τιμή είναι 1.81 (πρώτη) και η μεγαλύτερη 1.85 (τελευταία). Αφαιρώντας δύο τιμές από κάθε άκρη μένει η μεσαία μέτρηση, που είναι 1.83. Αυτή είναι η διάμεσος.

Η ελάχιστη, μέγιστη, μέση, και διάμεσος τιμή υπολογίζονται στο R απευθείας με τις συναρτήσεις `min, max, mean` και `median`, αντίστοιχα.

Αν κόψουμε το κάθε μισό στη μέση μπορούμε να βρούμε τη διάμεσο του κάθε μισού, που χωρίζουν το πρώτο τέταρτο και το τελευταίο τέταρτο των δεδομένων. Τα σημεία αυτά ονομάζονται *τεταρτημόρια*: Το πρώτο τεταρτημόριο (1st quartile) χωρίζει το χαμηλότερο 25%. Το δεύτερο τεταρτημόριο είναι η διάμεσος και χωρίζει το 50%. Το τρίτο τεταρτημόριο (3rd quartile) χωρίζει το υψηλότερο 25%. Αυτές είναι οι επιπλέον τιμές που μας δίνει η περίληψη του R. Βέβαια για τόσο λίγες τιμές που έχουμε εδώ αυτό δεν έχει πολύ νόημα, είναι όμως πάρα πολύ χρήσιμο σε μεταβλητές με δεκάδες ή εκατοντάδες μετρήσεις.

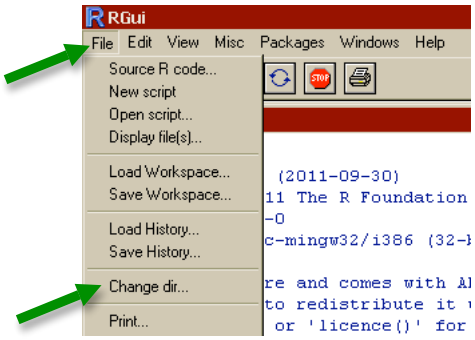
Η περίληψη εφαρμόζεται και σε ολόκληρα πλαίσια δεδομένων. Στην περίπτωση αυτή μας δίνει πληροφορίες για όλες τις μεταβλητές που περιλαμβάνονται στο πλαίσιο δεδομένων και προσαρμόζεται αυτόματα σε κάθε μεταβλητή αν είναι κατηγορική ή ποσοτική. Παράδειγμα:

```
> summary(atoma)
 sex        age
 F:1   Min.   :21.00
 M:1   1st Qu.:21.25
       Median :21.50
       Mean   :21.50
       3rd Qu.:21.75
       Max.   :22.00
```

Χρήση εξωτερικών αρχείων

Το R μπορεί να διαβάσει δεδομένα που έχουμε αποθηκευμένα σε αρχεία στο δίσκο του υπολογιστή μας. Με το R μπορούμε επίσης να αποθηκεύσουμε δεδομένα, αποτελέσματα επεξεργασίας, ή και τις εντολές και συναρτήσεις που χρησιμοποιήσαμε για την ανάλυσή μας.

Για να μπορεί να χρησιμοποιηθεί κάποιο εξωτερικό αρχείο πρέπει προηγουμένως να υποδείξουμε στο R σε ποιο φάκελο βρίσκονται τα αρχεία μας. Η επιλογή φακέλου γίνεται μέσα από τον κατάλογο επιλογών File → Change dir… (dir=directory, δηλαδή κατάλογος αρχείων). Με την επιλογή αυτή το R μας εμφανίζει το γνωστό παράθυρο επιλογής φακέλου των windows. Εντοπίζουμε και επιλέγουμε την τοποθεσία όπου βρίσκονται τα αρχεία μας.



Αφού επιλέξουμε τη σωστή τοποθεσία, μπορούμε να φορτώσουμε ένα πλαίσιο δεδομένων απευθείας από το δίσκο με τη συνάρτηση `read.table`, αναθέτοντας το περιεχόμενο απευθείας σε μια μεταβλητή. Π.χ., για να χρησιμοποιήσουμε τα κατηγορικά δεδομένα του 3ου κεφαλαίου του βιβλίου, τα αναθέτουμε στη μεταβλητή `ch3` ως εξής:

```
> read.table("chapter3_1.Rdata") -> ch31
```

Προσοχή, να μην ξεχνάμε την τελίτσα μέσα στο όνομα της συνάρτησης, χωρίς κενά!
Η συνάρτηση `str` μας δείχνει το αποτέλεσμα της ανάθεσης:

```
> str(ch31)
'data.frame':    264 obs. of  1 variable:
 $ education: Factor w/ 5 levels "Άλλο","Λύκειο",..: 5 5 5 5 5 5
5 5 5 5 ...
```

Πρόκειται για ένα πλαίσιο δεδομένων με μια μοναδική κατηγορική μεταβλητή με όνομα `education` η οποία περιέχει δεδομένα πέντε κατηγοριών. Τα στοιχεία των πρώτων σειρών ανήκουν όλα στην 5η κατηγορία.

Αργότερα θα δούμε πώς μπορούμε να αποθηκεύσουμε δικά μας δεδομένα καθώς και να χρησιμοποιήσουμε αρχεία αναλύσεων και εξωτερικά πακέτα συναρτήσεων.

Αθανάσιος Πρωτόπαπας
Μάρτιος 2012