ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

Πρόγραμμα Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Ανάλυση Δεδομένων με τη Χρήση Στατιστικών Πακέτων

Εισαγωγή στο R

Σημειώσεις παραδόσεων

Δημήτριος Αντζουλάκος

Πειραιάς 2013

(Α Έκδοση 2008)

Περιεχόμενα

5 Στοιχεία Πιθανοτήτων – Προσομοίωση

6 Θέματα Στατιστικής με το R

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Τι είναι το R

Το R είναι ένα τρία πράγματα: ένα έργο, μια γλώσσα και ένα περιβάλλον. Το R προσφέρει ένα ολοκληρωμένο σύνολο υπηρεσιών λογισμικού για ανάλυση δεδομένων, υπολογισμών και γραφημάτων. Το λογισμικό R γράφτηκε αρχικά από τους **R**oss Ihaka και **R**obert Gentleman στα μέσα της δεκαετίας του 90 (σε αυτούς οφείλει το όνομά του). Από το 1997 αναπτύσσεται από το R Development Core Team. Το R είναι λογισμικό ανοικτού κώδικα (open source) και αποτελεί μέρος του έργου GNU[1].

Σαν γλώσσα το R μπορεί να θεωρηθεί ότι αποτελεί μια εφαρμογή της γλώσσας προγραμματισμού S. Η γλώσσα και το περιβάλλον S αναπτύχθηκε στα Bell Laboratories (πρώην AT&T, τώρα Lucent Technologies) από τους Rick Becker, John Champers και Allan Wilks (το έργο ξεκίνησε το 1976). Άλλες δύο (εμπορικές) εφαρμογές της γλώσσας S είναι η "παλαιά S μηχανή" (S version 3; S-PLUS 3.x and 4.x) και η "νέα S μηχανή" (S version 4; S-PLUS 5.x και άνω). Όταν ρωτάμε για διαφορές μεταξύ R και S ουσιαστικά αναφερόμαστε στις διαφορές μεταξύ R και των δύο S μηχανών. Οι διαφορές είναι ελάχιστες και έτσι κώδικας που γράφεται για το R τρέχει σχεδόν αμετάβλητος και στις δύο S μηχανές.

Η ιστοσελίδα του R είναι η http://www.r-project.org και αποτελεί την κύρια πηγή πληροφόρησής του.

1.2 Εγκατάσταση

Για να εγκαταστήσετε το R πρέπει να επισκεφτείτε κάποιο mirror του CRAN[2] από την ιστοσελίδα http://cran.r-project.org. Επισκεφτείτε τη διεύθυνση http://cran.r-project.org/ και στο πλαίσιο Download and Install R που εμφανίζεται επιλέξτε • Download R for Windows .

[1] Το έργο GNU ξεκίνησε το 1984 με σκοπό την ανάπτυξη ενός ολοκληρωμένου λειτουργικού συστήματος (παρόμοιο με το Unix) το οποίο θα είναι ελεύθερο λογισμικό: το σύστημα GNU. Παραλλαγές του λειτουργικού συστήματος GNU που χρησιμοποιούν τον πυρήνα Linux χρησιμοποιούνται ευρέως στις μέρες μας. Αν και τα συστήματα αυτά αναφέρονται συνήθως με το όνομα "Linux", είναι πιο ακριβές να αποκαλούνται συστήματα GNU/Linux. Η λέξη GNU είναι αναδρομικό ακρωνύμιο του "GNU's Not Unix".

[2] Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for MacOS X
- Download R for Windows

Στην οθόνη που εμφανίζεται επιλέξτε base.

R for Windows

Subdirectories:

base	Binaries for base distribution (managed by Duncan Murdoch). This is what you want if you **install R for the first time**.
contrib	Binaries of contributed packages (managed by Uwe Ligges)

Η οθόνη που εμφανίζεται έχει ως τίτλο την τρέχουσα έκδοση της γλώσσας R (στην προκειμένη περίπτωση η 2.14.0). Επιλέξτε Download R 2.14.0 for Windows και σώστε το ομώνυμο αρχείο στην επιφάνεια εργασίας σας (Desktop).

R-3.0.2 for Windows (32/64 bit)

Download R 3.0.2 for Windows (52 megabytes, 32/64 bit)

Installation and other instructions

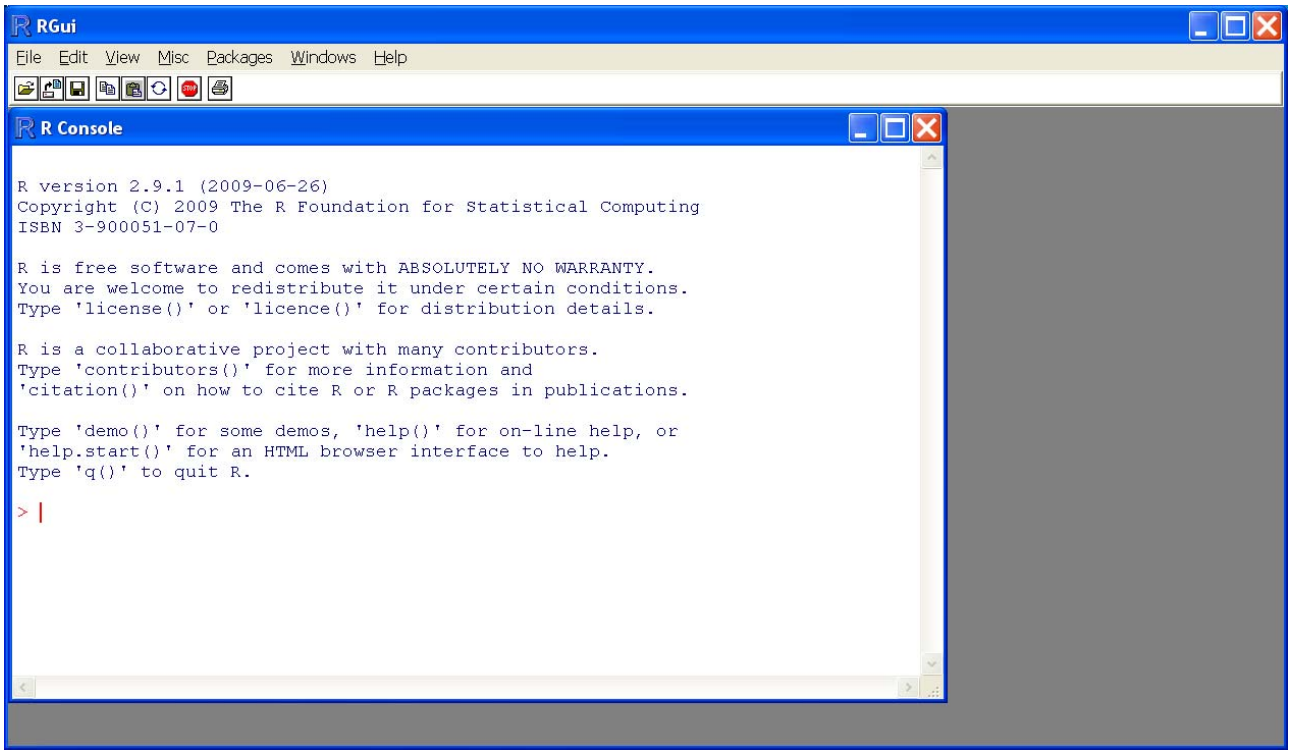
New features in this version

Τρέξτε το αρχείο R-2.14.0-win.exe από την επιφάνεια εργασίας σας κάνοντας διπλό κλικ πάνω του διατηρώντας τις προεπιλεγμένες επιλογές που εμφανίζονται κατά τη διάρκεια εγκατάστασης του προγράμματος (στο παράθυρο Select Components τικάρετε όλες τις επιλογές). Το πρόγραμμα εγκατάστασης δημιουργεί τον κατάλογο C:\Program Files\R\R-2.14.0 που η ονομασία του ποικίλει ανάλογα με την έκδοση του R που εγκαθιστάτε (στην περίπτωσή μας η 2.14.0). Το R πρόγραμμα βρίσκεται στον κατάλογο C:\Program Files\R\R-2.14.0\bin\i386 και φέρει το όνομα Rgui.exe[3]. Επίσης δημιουργείται εικονίδιο συντόμευσης στην επιφάνεια εργασίας και ενημερώνεται και το μενού εκκίνησης των Windows.

[3] GUI: Graphical User Interface

1.3 Εκκίνηση, παράθυρα και τερματισμός του R

Ο πιο απλός τρόπος να θέσουμε σε εκκίνηση το R είναι να κάνουμε διπλό κλικ στο εικονίδιο συντόμευσής του που βρίσκεται στην επιφάνεια εργασίας. Τότε εμφανίζεται το κυρίως παράθυρο της εφαρμογής RGui με 7 μενού, μια γραμμή εργαλείων και το παράθυρο της R κονσόλας (R console). Στην R κονσόλα δίνονται κάποιες εισαγωγικές πληροφορίες και στο τέλος εμφανίζεται το σύμβολο υποβολής εντολών ">" σε κόκκινο χρώμα.

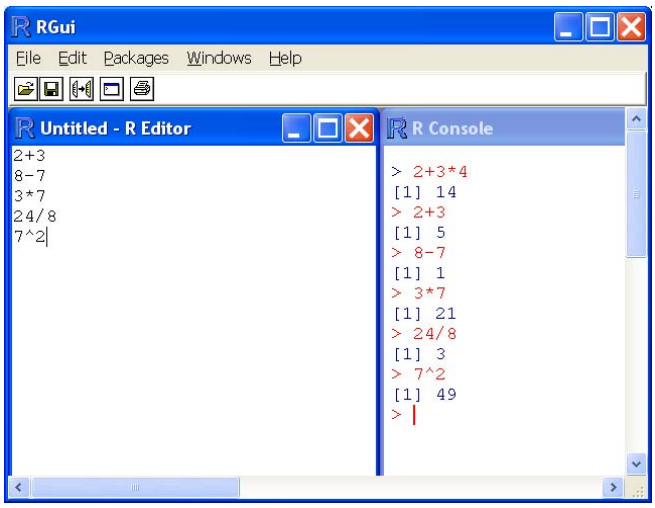


Μπορούμε να διαγράψουμε το περιεχόμενο της R κονσόλας επιλέγοντας Edit>Clear Console.

Για να εκτελεστεί μια εντολή την πληκτρολογούμε και μετά πατάμε το πλήκτρο Enter. Για παράδειγμα

```
> 2+3*4
[1] 14
```

Αν η εντολή είναι μεγάλη και δεν χωρά σε μια γραμμή ή όταν καταχωρηθεί μια μη πλήρης εντολή εμφανίζεται το σύμβολο "+" στην επόμενη γραμμή για να καταχωρηθεί το υπόλοιπο μέρος της εντολής. Μια γραμμή μπορεί να έχει μέχρι 128 χαρακτήρες.



Όταν ο όγκος των εντολών είναι μεγάλος τότε συμφέρει να εισάγονται με χρήση του παραθύρου του R συντάκτη (R Editor) επιλέγοντας

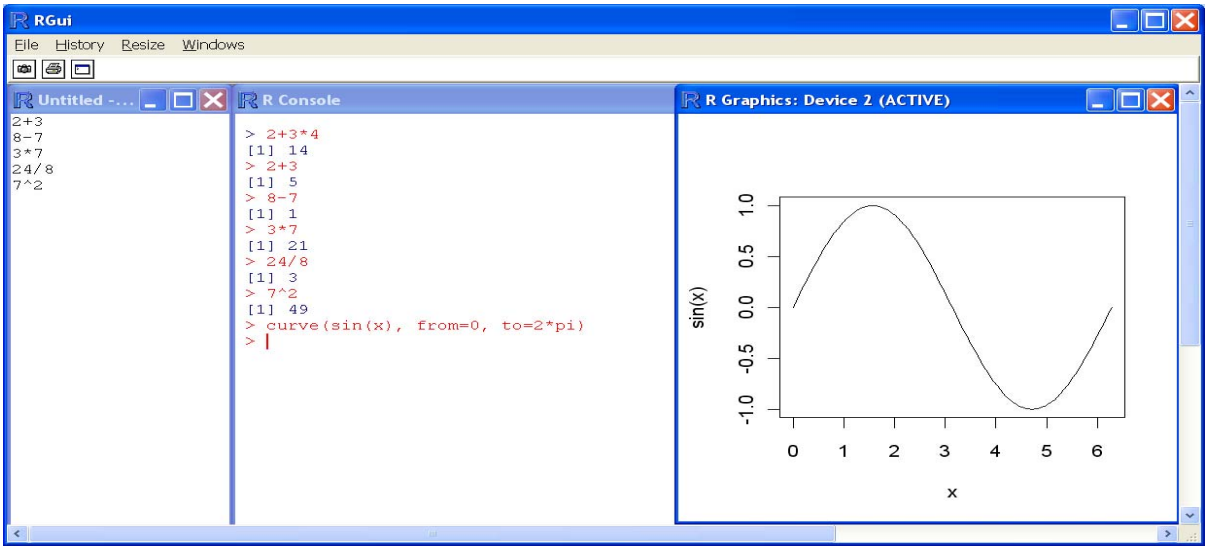
File>New script ... . Το πλεονέκτημα του R συντάκτη είναι ότι μπορούμε να γράψουμε όσες εντολές θέλουμε και μετά να εκτελέσουμε όσες από αυτές θέλουμε αρκεί να τις επιλέξουμε (μαρκάρουμε) και να πατήσουμε "Ctrl+R". Εναλλακτικά μπορούμε να χρησιμοποιήσουμε το μενού Edit (Run line or selection, Run all). Τα αποτελέσματα που προκύπτουν εμφανίζονται στην R κονσόλα.

Μπορούμε να εμφανίζουμε στην R κονσόλα μπροστά από το σύμβολο υποβολής εντολών ">" όλες τις εντολές που έχουν εκτελεστεί στην R κονσόλα με τα βελάκια "↑" και "↓". Επίσης με τα βελάκια "→" και "←" μπορούμε να μετακινηθούμε μέσα σε μια εντολή. Πατώντας το πλήκτρο Home (End) του πληκτρολογίου μας μεταφερόμαστε στην αρχή (τέλος) της γραμμής που βρισκόμαστε. Με το πλήκτρο Bakspace ή το πλήκτρο Delete διαγράφουμε τους χαρακτήρες που βρίσκονται αριστερά ή δεξιά του κέρσορα, αντίστοιχα.

Αν δοθεί εντολή κατασκευής γραφήματος, το γράφημα εμφανίζεται σε ξεχωριστό παράθυρο (R γραφικά - R Graphics). Για παράδειγμα, εκτελώντας την εντολή

```
> curve(sin(x), from=0, to=2*pi)
```

εμφανίζεται η γραφική παράσταση της συνάρτησης $\sin x$ στο διάστημα $[0, 2\pi]$.



Για να τερματίσουμε το R πληκτρολογούμε

```
> q()
```

(συντομογραφία της συνάρτησης `quit`) στην R κονσόλα, ή επιλέγουμε File>Exit.

1.4 Βοήθεια

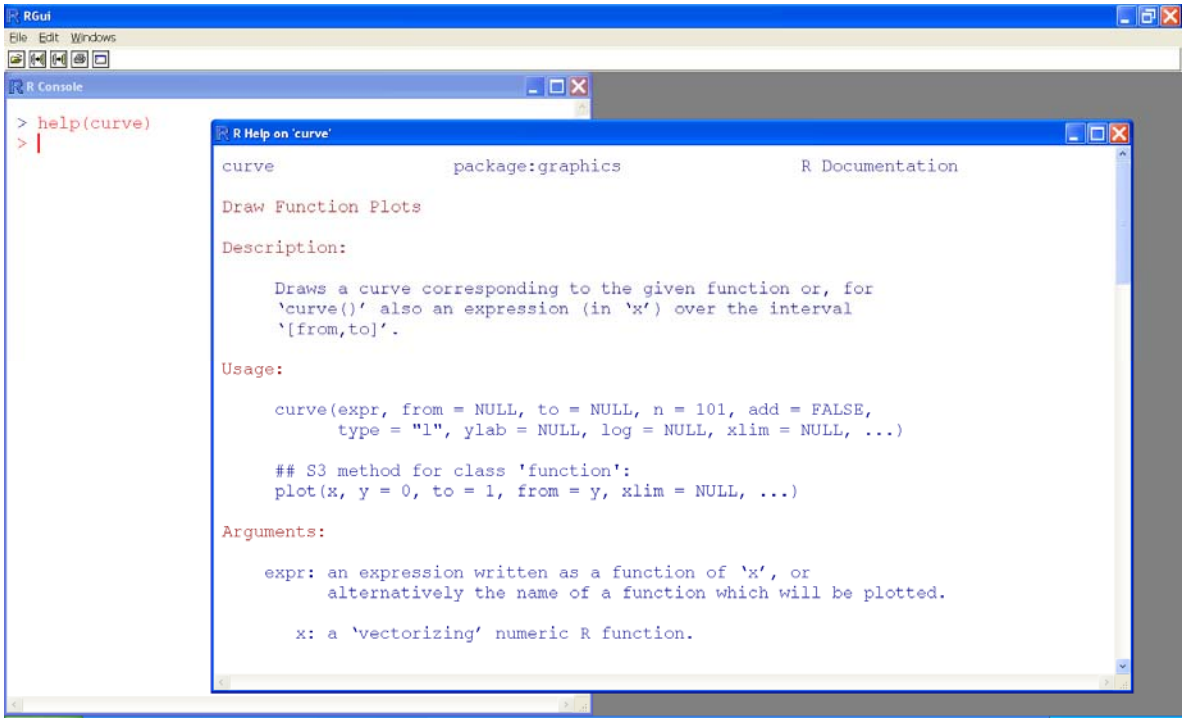
Για να λάβουμε βοήθεια για τη σύνταξη και τα ορίσματα μιας συνάρτησης του R χρησιμοποιούμε τη συνάρτηση `help(όνομα συνάρτησης)` ή εναλλακτικά `?όνομα συνάρτησης`. Συνεπώς εκτελώντας την εντολή

```
> help(curve)
```

ή εναλλακτικά

```
> ?curve
```

εμφανίζεται η περιγραφή, η χρήση, τα ορίσματα, λεπτομέρειες καθώς επίσης και παραδείγματα της συνάρτησης `curve` του πακέτου graphics (δείτε Παράγραφο 1.6 για την έννοια του πακέτου).

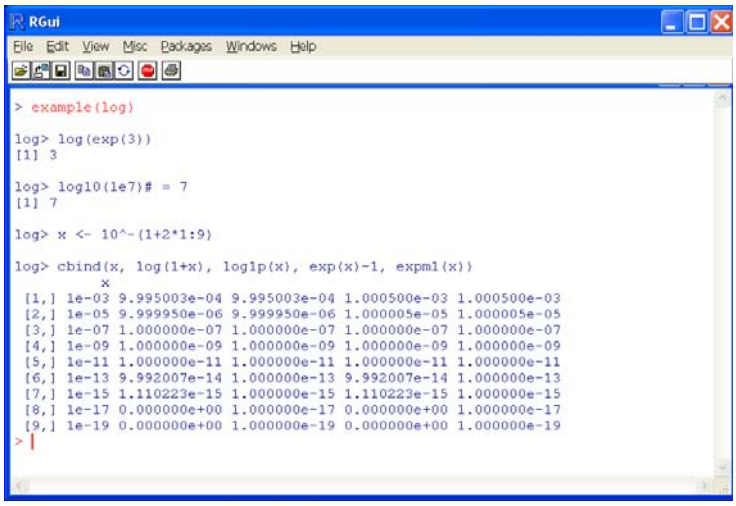


Αν δεν θυμόμαστε το πλήρες όνομα της συνάρτησης που θέλουμε να εξετάσουμε, αλλά ένα μέρος του ονόματός της, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση `apropos("μέρος ονόματος συνάρτησης")` για να λάβουμε μια λίστα με συναρτήσεις που περιέχουν στην ονομασία τους το συγκεκριμένο μέρος του ονόματος. Για παράδειγμα

```
> apropos("cu")
 [1] "cummax"              "cummin"              "cumprod"
 [4] "cumsum"              "curve"               "cut"
 [7] "cut.Date"            "cut.default"         "cut.POSIXt"
[10] "cutree"              "dev.cur"             "icuSetCollate"
[13] "is.recursive"        "occupationalStatus"
```

Η συνάρτηση `help.start` ξεκινά βοήθεια για το R σε web μορφή. Επίσης, ολοκληρωμένη βοήθεια για το R είναι προσβάσιμη και μέσω του μενού Help της κύριας εφαρμογής RGui. Για να εμφανιστεί βοήθεια για τη συνάρτηση `curve` ακολουθούμε τη διαδοχή Help>R functions(text)... , στο παράθυρο διαλόγου που εμφανίζεται πληκτρολογούμε curve στο πλαίσιο Help on, και πατάμε OK. Επίσης και η εκτέλεση της συνάρτησης `apropos` μπορεί να γίνει μέσω της διαδρομής Help>Apropos ... .

Για να εμφανιστούν παραδείγματα εφαρμογής μιας συνάρτησης (ή θέματος) χρησιμοποιούμε τη συνάρτηση `example(όνομα συνάρτησης)`. Για παράδειγμα εκτελώντας την εντολή `example(log)` προκύπτει το διπλανό πλαίσιο. Σημειώνεται ότι ο αριθμός `1.5e3` είναι ίσος με $1.5\times10^{3}=1500$, ενώ ο αριθμός `1.5e-3` είναι ίσος με $1.5\times10^{-3}=0.0015$.



Μια πλήρης σειρά εγχειριδίων για το R σε μορφή *.pdf αρχείων βρίσκεται στην ιστοσελίδα http://cran.r-project.org/ (Manuals). Ιδιαίτερη αναφορά αξίζει στο εγχειρίδιο με τίτλο "The R Reference Index" (είναι περίπου 3500 σελίδες).

1.5 Βασικά στοιχεία σύνταξης εντολών

Η σύνταξη εκφράσεων στη γλώσσα R γίνεται χρησιμοποιώντας αλφαριθμητικές εκφράσεις. Το σύνολο των χαρακτήρων που μπορούν να χρησιμοποιηθούν στις εκφράσεις εξαρτάται από το λειτουργικό σύστημα και τη χώρα εντός της οποίας εκτελείται το R. Μπορούν επίσης να χρησιμοποιηθούν τα σύμβολα "`.`" και "`_`" . Ένα όνομα μπορεί να ξεκινά με "`.`" ή με ένα γράμμα, αλλά αν ξεκινά με "`.`" ο δεύτερος χαρακτήρας δεν μπορεί να είναι αριθμός. Υπάρχει διάκριση πεζών και κεφαλαίων γραμμάτων (case sensitive) και επομένως το b και το B είναι διαφορετικά σύμβολα.

Οι στοιχειώδεις εντολές (commands) μπορεί να είναι είτε εκφράσεις (expressions), είτε εκχωρήσεις (assignments). Οι εκφράσεις υπολογίζονται, εμφανίζονται στην οθόνη και η τιμή τους στη συνέχεια χάνεται. Για παράδειγμα

```
> exp(1)+1
[1] 3.718282
```

Μια εντολή εκχώρησης δύναται να υπολογίσει μια έκφραση, καταχωρεί το αποτέλεσμα σε μια μεταβλητή (αντικείμενο – object) με τη βοήθεια του συμβόλου εκχώρησης "`<-`" (ή ισοδύναμα με το "`=`"), αλλά δεν το εμφανίζει στην οθόνη εκτός και αν ζητηθεί. Για παράδειγμα

```
> x <- 3+4
> x
[1] 7
```

Κάθε εντολή καταχωρείται συνήθως σε ξεχωριστή γραμμή. Μπορούμε να καταχωρήσουμε αρκετές εντολές σε μια γραμμή αρκεί να διαχωρίζονται με το σύμβολο "`;`". Για παράδειγμα

```
> sqrt(81);y <- log(10, base=exp(1));y;z <- exp(y);z
[1] 9
[1] 2.302585
[1] 10
```

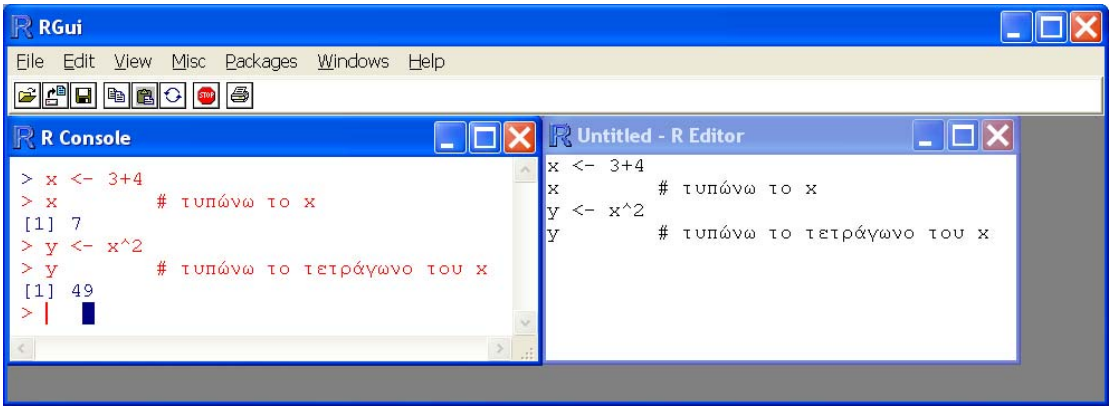
Στοιχειώδης εντολές μπορούν να ομαδοποιηθούν σε μια σύνθετη εντολή αν περιβληθούν με άγκιστρα. Για παράδειγμα

```
> {w <-5;w;w+2}
[1] 7
```

Προσέξτε τη διαφορά με την ακόλουθη σύνθετη εντολή

```
{w <-5;print(w);w+2}
[1] 5
[1] 7
```

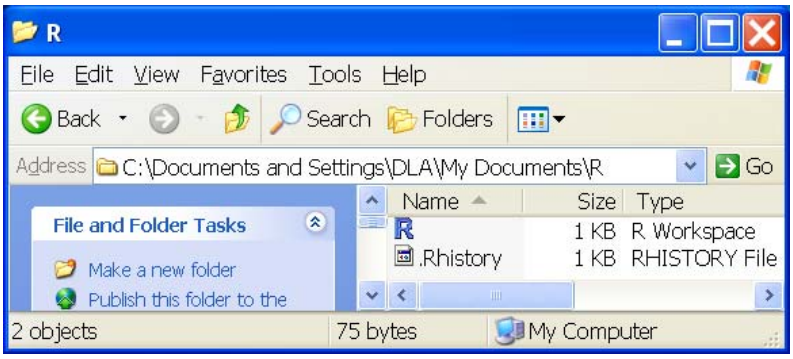
Ότι γραφεί σε μια γραμμή μετά το σύμβολο "#" θεωρείται σχόλιο. Τα σχόλια είναι χρήσιμα κυρίως όταν χρησιμοποιούμε τον R συντάκτη.



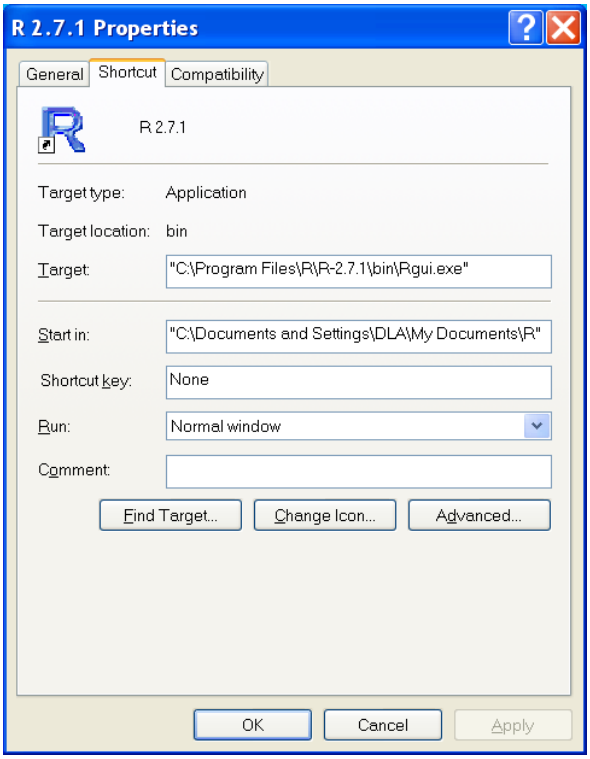
Οι εντολές που μπορούν να εισαχθούν στην R κονσόλα δεν μπορούν να υπερβαίνουν σε χωρητικότητα τα 1024 bytes.

1.6 Αποθήκευση

Όταν ξεκινά το R ψάχνει να βρει στον κατάλογο εργασίας (working directory) το αρχείο *.RData (περιέχει objects) και το αρχείο *.Rhistory (περιέχει εντολές) για να τα φορτώσει αυτόματα. Για να βρούμε τον κατάλογο εργασίας του R κάνουμε δεξί κλικ πάνω στο εικονίδιο συντόμευσης του R στην επιφάνεια εργασίας, επιλέγουμε Properties, κάνουμε κλικ στο μενού



Shortcut, και ο κατάλογος εργασίας εμφανίζεται στο πεδίο Start in. Ο κατάλογος εργασίας μπορεί να αλλάξει. Στο διπλανό σχήμα είναι ο C:\Documents and Settings\DLA\My Documents\R (ο προεπιλεγμένος κατάλογος εργασίας είναι ο C:\Documents and Settings\ ... \My Documents).

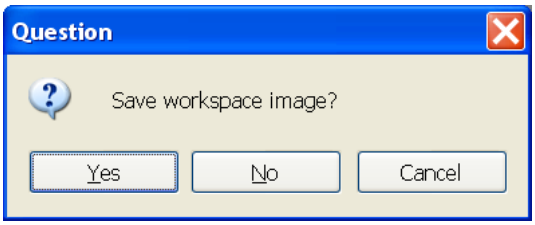


Ο όρος αντικείμενα (objects) χρησιμοποιείται για να δηλώσει τις "οντότητες" που δημιουργούνται και χειρίζονται στο R. Αντικείμενα είναι μεταβλητές, διανύσματα αριθμών, μια σειρά χαρακτήρων, συναρτήσεις και γενικότερα σύνθετες δομές που χτίζονται από αυτά τα στοιχεία.

Μόλις γίνει εκκίνηση του R μπορούμε να εμφανίσουμε στην R κονσόλα τις αποθηκευμένες εντολές που βρίσκονται στο αρχείο *.Rhistory δουλεύοντας με τα βελάκια "↑" και "↓". Τα αποθηκευμένα αντικείμενα που βρίσκονται στο αρχείο *.RData εμφανίζονται με τη συνάρτηση `objects` (εναλλακτικά μπορεί να χρησιμοποιηθεί η διαδρομή Misc>List objects). Για παράδειγμα

```
> objects()
character(0)
```

που σημαίνει ότι δεν υπάρχουν αποθηκευμένα αντικείμενα στον κατάλογο εργασίας μας στο αρχείο *.RData (ή ότι δεν υπάρχει καν αρχείο *.RData). Για να αποθηκεύσουμε στον κατάλογο εργασίας μας τις εντολές που υπάρχουν στην R κονσόλα, χωρίς να τερματίσουμε το R, χρησιμοποιούμε τη συνάρτηση `savehistory` που δημιουργεί το αρχείο *.Rhistory. Για να αποθηκεύσουμε στον κατάλογο εργασίας μας τα αντικείμενα που υπάρχουν στην R κονσόλα, χωρίς να τερματίσουμε το R, χρησιμοποιούμε τη συνάρτηση `save.image` που δημιουργεί το αρχείο *.RData. Για να δημιουργηθούν και τα δύο αρχεία όταν αποφασίσουμε να τερματίσουμε το R ακολουθούμε τη διαδοχή File>Exit και επιλέγουμε Yes στο παράθυρο Question. Σημειώνουμε ότι μπορούμε να επιλέξουμε τα ονόματα των αρχείων που περιέχουν τις εντολές και τα αντικείμενα εκτελώντας, αντίστοιχα, τις εντολές `savehistory(file="όνομα.Rhistory")`, `save.image(file="όνομα.RData")`.



Ο πιο απλός τρόπος για να φορτωθούν τα δύο παραπάνω αρχεία την επόμενη φορά που θα εκκινή-

σουμε το R (δεν φορτώνονται αυτόματα) είναι μέσω της διαδρομής File>Load History ... και File>Load Workspace ..., αντίστοιχα.

Αξίζει να σημειώσουμε ότι τα αρχεία που περιέχουν εντολές (*.Rhistory) μπορούμε να τα επεξεργαστούμε με οποιοδήποτε επεξεργαστή κειμένου (Word, Notepad, κτλ.). Επίσης μπορούμε να αποθηκεύσουμε οτιδήποτε εμφανίζεται στην R κονσόλα σε ένα αρχείο *.txt ακολουθώντας τη διαδρομή File>Save to File ... .

Για να σώσουμε το περιεχόμενα του R συντάκτη πατάμε "Ctrl+S" και σώζουμε τα περιεχόμενα σε ένα .R αρχείο. Εναλλακτικά μπορούμε να χρησιμοποιήσουμε το μενού File>Save as... . Το .R αρχείο φορτώνεται μελλοντικά με τη διαδρομή File>Open script… .

Σημειώνεται ότι παρόμοια δουλειά με τη συνάρτηση `objects` κάνει και η συνάρτηση `ls`. Για να διαγράψουμε κάποια αντικείμενα χρησιμοποιούμε τη συνάρτηση `rm(όνομα πρώτου αντικειμένου, ..., όνομα_τελευταίου_αντικειμένου)`. Εναλλακτικά μπορεί να χρησιμοποιηθεί η διαδρομή Misc>Remove all objects για τη διαγραφή όλων των αντικειμένων ή εκτέλεση της εντολής `rm(list=ls())`. Για παράδειγμα

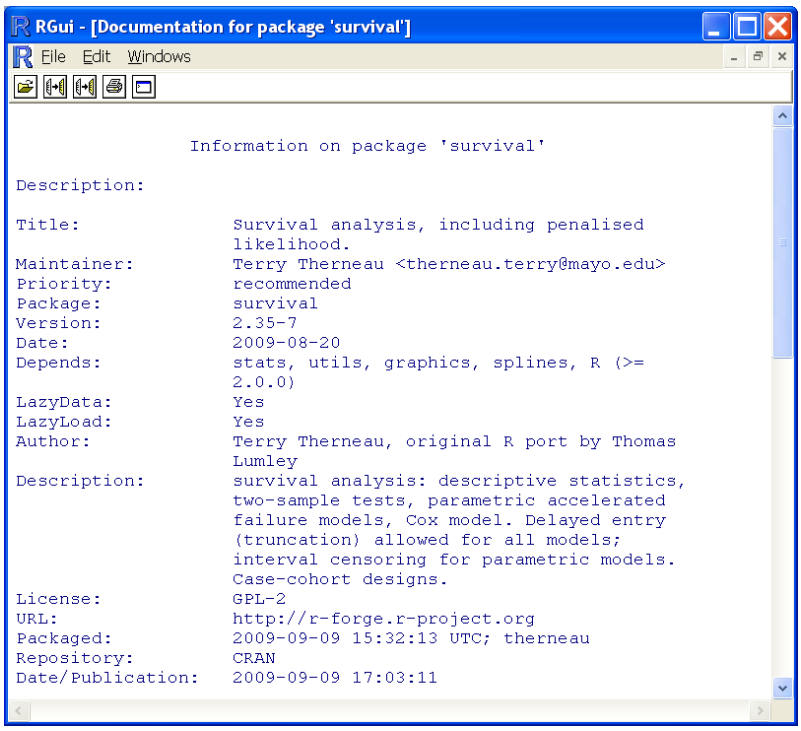
```
> x<-5;y<-7;z<-9
> ls()
[1] "x" "y" "z"
> rm(x,y)
> ls()
[1] "z"
```

1.7 Πακέτα

Όλες οι συναρτήσεις (functions) και τα σύνολα δεδομένων (datasets) που χρησιμοποιεί το R είναι αποθηκευμένα σε πακέτα (packages). Οι συναρτήσεις και τα σύνολα δεδομένων ενός πακέτου είναι διαθέσιμα για χρήση μόνο όταν φορτωθεί το πακέτο. Για να δούμε ποια πακέτα είναι εγκατεστημένα χρησιμοποιούμε τη συνάρτηση `library` και για να δούμε ποια έχουν φορτωθεί χρησιμοποιούμε τη συνάρτηση `search`.

Για να πάρουμε πληροφορίες για ένα πακέτο εκτελούμε την εντολή `library(help=όνομα πακέτου)`. Για παράδειγμα

```
> library(help=survival)
```



Για να φορτώσουμε ένα πακέτο χρησιμοποιούμε τη συνάρτηση `library(όνομα_πακέτου)` ή εναλλακτικά μπορούμε να χρησιμοποιήσουμε τη διαδρομή Packages>Load package ... .

Εκτός από τα βασικά πακέτα του R, υπάρχουν διαθέσιμα εκατοντάδες πακέτα που καλύπτουν σχεδόν όλες τις περιοχές της στατιστικής. Τα πακέτα αυτά μπορούν να εγκατασταθούν (και στη συνέχεια να φορτωθούν) ακολουθώντας τη διαδρομή Packages>Install package(s) ... . Ο κατάλογος των πακέτων που διατίθενται για εγκατάσταση δίνεται στην ιστοσελίδα http://cran.r-project.org επιλέγοντας ένα Mirrors στο μενού CRAN και μετά Packages στο μενού Software. Για παράδειγμα, δύο πακέτα που σχετίζονται με το Στατιστικό Έλεγχο Ποιότητας είναι τα qcc (Quality Control Charts) και spc (Statistical Process Conrol). Για κάθε πακέτο υπάρχει εγχειρίδιο αναφοράς (reference manual) σε μορφή *.pdf αλλά και το ίδιο το πακέτο σε μορφή *.zip σε περίπτωση που θέλουμε να το σώσουμε σε ένα κατάλογο (directory) και κατόπιν να το εγκαταστήσουμε χρησιμοποιώντας τη διαδρομή Packages>Install package(s) from local zip files.

qcc: Quality Control Charts

Shewhart quality control charts for continuous, attribute and count data. Cusum and EWMA charts. Operating characteristic curves. Process capability analysis. Pareto chart and cause-and-effect chart.

Version:	1.3
Depends:	R (? 2.6)
Published:	2008-10-12
Author:	Luca Scrucca
Maintainer:	Luca Scrucca <luca at stat.unipg.it>
License:	GPL (? 2)
Citation:	qcc citation info
CRAN checks:	qcc results

Downloads:

Package source:	qcc_1.3.tar.gz
MacOS X binary:	qcc_1.3.tgz
Windows binary:	qcc_1.3.zip
Reference manual:	qcc.pdf
Old sources:	qcc archive

Επίσης τα εγκατεστημένα πακέτα μπορούν να ενημερωθούν (update) όποια στιγμή θέλουμε χρησιμοποιώντας τη διαδρομή Packages>Update packages ....

1.8 Βασικές συναρτήσεις και λογικοί τελεστές

Ο ακόλουθος πίνακας δίνει τους αριθμητικούς τελεστές που χρησιμοποιούνται σε βασικές αριθμητικές πράξεις

Πίνακας 1.1: Αριθμητικοί τελεστές

Τελεστής	Πράξη
+	Πρόσθεση
-	Αφαίρεση
*	Πολλαπλασιασμός
/	Διαίρεση
^	Ύψωση σε δύναμη
%/%	Ακέραια μέρη διαίρεσης
%%	Υπόλοιπο διαίρεσης

Ο ακόλουθος πίνακας δίνει λογικούς τελεστές και τελεστές σχέσεων

Πίνακας 1.2: Λογικοί τελεστές και τελεστές σχέσεων

Τελεστής	Ερμηνεία
&	και
|	ή
!	όχι
==	ίσο με
!=	άνισο από
<	μικρότερο από
<=	μικρότερο ή ίσο από
>	μεγαλύτερο από
>=	μεγαλύτερο ή ίσο από

Ο ακόλουθος πίνακας δίνει ορισμένες γνωστές συναρτήσεις του R

Πίνακας 1.3: Βασικές συναρτήσεις

Συνάρτηση	Λειτουργία
`sqrt(x)`	Τετραγωνική ρίζα
`abs(x)`	Απόλυτη τιμή (ή μέτρο)
`sin(x), cos(x), tan(x)`	Τριγωνομετρικές συναρτήσεις
`asin(x), acos(x), atan(x)`	Τόξα τριγωνομετρικών συναρτήσεων
`factorial(x)`	Παραγοντικό

`choose(n,x)`	Διωνυμικός συντελεστής (n ανά x)
`exp(x)`	Εκθετική συνάρτηση
`log(x)`	Λογάριθμος (φυσικός)
`logb(x),log(x,b)`	Λογάριθμος με βάση το b
`gamma(x)`	Συνάρτηση Γάμμα
`floor(x)`	Μικρότερος ακέραιος $\geq x$
`ceiling(x)`	Μεγαλύτερος ακέραιος $\leq x$
`round(x, digits=n)`	Στρογγυλοποίηση
`signif(x, digits=6)`	Σημαντικά ψηφία

1.9 Ένα παράδειγμα

Κλείνοντας το πρώτο εισαγωγικό κεφάλαιο δίνουμε το ακόλουθο παράδειγμα επίδειξης υπολογισμών και συναρτήσεων του R.

```
> 4/0;-3/0                                  # + και - άπειρο
[1] Inf
[1] -Inf
> 0/Inf;0/0;exp(-Inf);Inf-Inf               # NaN (Not a number)
[1] 0
[1] NaN
[1] 0
[1] NaN
> abs(-3.1);abs(5.7)                        # Απόλυτη τιμή αριθμού
[1] 3.1
[1] 5.7
> sqrt(81); 81^(1/2)                        # Τετραγωνική ρίζα
[1] 9
[1] 9
> exp(1);exp(2)                             # Εκθετική συνάρτηση
[1] 2.718282
[1] 7.389056
> log(exp(1));log(20)                       # Φυσικός λογάριθμος
[1] 1
[1] 2.995732
> log2(64);log10(1000)                      # Λογάριθμος με άλλες βάσεις I
[1] 6
[1] 3
> log(64,2);log(1000,10)                    # Λογάριθμος με άλλες βάσεις II
[1] 6
[1] 3
>
> floor(3.7);floor(-3.7);floor(3)           # Μικρότερος ακέραιος
[1] 3
[1] -4
[1] 3
> ceiling(3.7);ceiling(-3.7);ceiling(3)     # Μεγαλύτερος ακέραιος
[1] 4
[1] -3
```

```
[1] 3
> round(exp(1), digits=2)                # Στρογγυλοποίηση
[1] 2.72
> signif(exp(1), digits=6)               # Σημαντικά ψηφία
[1] 2.71828
> 121%/%7; 121%%7                     # Ακέραια μέρη διαίρεσης και υπόλοιπο
[1] 17
[1] 2
> sin(pi/6);cos(pi/6);tan(pi/6)          # Τριγωνομετρικές συναρτήσεις
[1] 0.5
[1] 0.8660254
[1] 0.5773503
> asin(sqrt(3)/2);pi/3;atan(1);pi/4      # Αντίστροφες τριγωνομετρικές
συναρτήσεις
[1] 1.047198
[1] 1.047198
[1] 0.7853982
[1] 0.7853982
> x<-10;x<4;x+2==12                   #  Λογικές συνθήκες
[1] FALSE
[1] TRUE
```

1.10 Σύνοψη εντολών Κεφαλαίου 1

```
abs, acos, apropos, asin, atan
ceiling, class, cos, curve, choose
example, exp
floor, factorial
gamma
help, help.start
library, log, log10, ls
objects
print
quit
rm, round
savehistory, save.image, search, signif, sin, sqrt
tan
```

ΚΕΦΑΛΑΙΟ 2

Αντικείμενα Δεδομένων

2.1 Βαθμωτά αντικείμενα (scalar objects)

Το πιο απλό είδος αντικειμένου στο R είναι τα βαθμωτά (scalar) αντικείμενα αριθμητικού τύπου[4] (numeric mode), δηλαδή αντικείμενα με μία μόνο αριθμητική τιμή που ωστόσο θεωρούνται διανύσματα με μια συνιστώσα. Για παράδειγμα

```
> # Δημιουργία του βαθμωτού αντικειμένου x με τιμή 7
> x <- 7
> # Δημιουργία του βαθμωτού αντικειμένου y με τιμή 2
> y <- 2
```

(εναλλακτικά μπορεί να χρησιμοποιηθεί η συνάρτηση `assign` για να εκχωρήσουμε στη μεταβλητή *x* την τιμή 7, ως `assign("x", 7)`). Τέτοιου είδους αντικείμενα (μεταβλητές) συναντήσαμε στο Κεφάλαιο 1. Αυτά τα αντικείμενα επιδέχονται αλγεβρικούς χειρισμούς. Για παράδειγμα

```
> z <- x+y; z; x-y; x+y+z
[1] 9
[1] 5
[1] 18
```

Επίσης υπάρχουν βαθμωτά αντικείμενα λογικού (logical) τύπου, ή τύπου χαρακτήρα (character). Οι τιμές ενός λογικού αντικειμένου είναι TRUE, FALSE ή και NA (not available), και εκχωρούνται άμεσα με T, F ή NA, αντίστοιχα. Οι τιμές TRUE, FALSE προκύπτουν συνήθως ως τιμές λογικών εκφράσεων. Για παράδειγμα

```
> # Δημιουργία των βαθμωτών αντικειμένων a, b, c με τιμές TRUE, FALSE, NA
> a <- T; b <- F; c <- NA
> a; b; c
[1] TRUE
[1] FALSE
[1] NA
> # Δημιουργία των βαθμωτών αντικειμένων w, s με λογικές εκφράσεις
> w <- x>10; w; s <- y==2
[1] FALSE
[1] TRUE
> # Δημιουργία των βαθμωτών αντικειμένων k και m με τιμές kostas και
maria
> k <- "kostas"; m <- "maria"
> k; m
[1] "kostas"
[1] "maria"
```

[4] Οι τιμές της συνάρτησης `mode` για βαθμωτά αντικείμενα ή διανύσματα (vectors) είναι logical, numeric, complex, character ή raw. Για λίστες (Lists) είναι list.

Η συνάρτηση `mode` μας δίνει τον τύπο κάθε αντικειμένου ενώ η συνάρτηση `class` την τάξη[5] του. Για παράδειγμα

```
> mode(x); mode(c); mode(m)
[1] "numeric"
[1] "logical"
[1] "character"
> z <- 10+9i; mode(z)
[1] "complex"
```

2.2 Διανύσματα (vectors)

Ο πιο απλός τρόπος για να ορίσουμε το διάνυσμα $v1 = (1, 2, 3, 4)$ στο R είναι μέσω της συνάρτησης `c`[6]. Έτσι

```
> v1 <- c(1,2,3,4)
> v1
[1] 1 2 3 4
```

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε τη συνάρτηση `scan` για να εισάγουμε μια προς μια τα στοιχεία του διανύσματος. Για να ορίσουμε το διάνυσμα $v2 = (5, 6, 7)$ εκτελούμε την εντολή

```
> v2 <- scan()
```

και στη συνέχεια δίνουμε τα στοιχεία του διανύσματος ένα προς ένα πατώντας μετά από κάθε εισαγωγή το πλήκτρο Enter. Όταν ζητηθεί το $4^{ο}$ στοιχείο δεν εισάγουμε τίποτα αλλά πατάμε απλά ακόμη μια φορά το πλήκτρο Enter για να δηλωθεί το τέλος εισαγωγής των συνιστωσών του διανύσματος $v2$. Έτσι

```
> v2 <-scan()
1: 5
2: 6
3: 7
4:
Read 3 items
> v2
[1] 5 6 7
```

Τα βαθμωτά αντικείμενα, όπως προαναφέραμε, θεωρούνται διανύσματα με μια συνιστώσα

```
> x <- 3; length(x)
[1] 1
```

Με τα διανύσματα μπορούμε να κάνουμε πράξεις και να εφαρμόσουμε συναρτήσεις. Για παράδειγμα

```
>  v1/2;  exp(v1);  min(v1);  max(v1);  range(v2);  length(v2);  sum(v1);
prod(v1); mean(v2); var(v1)
[1] 0.5 1.0 1.5 2.0
```

[5] Οι τιμές της συνάρτησης class μπορεί να είναι numeric, integer, complex, logical, character, list, matrix, array, factor, ts, data.frame.

[6] Το γράμμα *c* προέρχεται από τη λέξη *concatenate* που σημαίνει συνδέω αλυσιδωτά (κρικώνω).

```
[1]  2.718282  7.389056 20.085537 54.598150
[1] 1
[1] 4
[1] 5 7
[1] 3
[1] 10
[1] 24
[1] 6
[1] 1.666667
```

(η συνάρτηση `var(x)` είναι ίση με `sum((x-mean(x))^2)/(length(x)-1)`, δηλαδή δίνει τη "δειγματική διακύμανση" του δείγματος με τιμές τα στοιχεία του διανύσματος *x*).

Όταν εκτελούμε πράξεις με διανύσματα αυτές γίνονται κατά συντεταγμένες. Αν δύο διανύσματα δεν είναι της ίδιας διάστασης τότε το διάνυσμα με τη μικρότερη διάσταση επαναλαμβάνει τα στοιχεία του όσες φορές χρειαστεί για να αποκτήσει τη διάσταση του διανύσματος με τη μεγαλύτερη διάσταση έτσι ώστε να είναι δυνατή η πραγματοποίηση πράξεων μεταξύ των διανυσμάτων. Για παράδειγμα

```
> v1+2*v2               # (1,2,3,4)+2*(5,6,7,5)
[1] 11 14 17 14
Warning message:
In v1 + 2 * v2 :
  longer object length is not a multiple of shorter object length

> v1/v2                 # (1/5,2/6,3/7,4/5)
[1] 0.2000000 0.3333333 0.4285714 0.8000000
Warning message:
In v1/v2 :
  longer object length is not a multiple of shorter object length
```

Με τη συνάρτηση `c` μπορούμε επίσης να "ενώσουμε" διανύσματα. Για παράδειγμα

```
> c(v1,v2,v2^2,2*v1)
[1]  1  2  3  4  5  6  7 25 36 49  2  4  6  8
> c(c(2,3), c(12,23,34))
[1]  2  3 12 23 34
```

Ειδικά διανύσματα μπορούν να οριστούν και με άλλους πιο απλούς τρόπους. Για παράδειγμα, `1:10` είναι το διάνυσμα $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$, ενώ το `10:1` (ή `rev(1:10)`) είναι το διάνυσμα $(10, 9, 8, 7, 6, 5, 4, 3, 2, 1)$. Ο τελεστής "`:`" έχει προτεραιότητα όταν εκτελούνται πράξεις. Για παράδειγμα

```
> v3 <- 2*1:20
> v3
[1]  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
```

Ειδικά διανύσματα μπορούν να δημιουργηθούν και με τη συνάρτηση `seq` που έχει επιπρόσθετες δυνατότητες. Για παράδειγμα

```
> seq(1,30)
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
[23] 23 24 25 26 27 28 29 30
> seq(-1,1,0.2)
[1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
> seq(2, by=0.5, length=12)
[1] 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5
> seq(2,10,length=4)
[1]  2.000000  4.666667  7.333333 10.000000
```

Μια άλλη χρήσιμη συνάρτηση για τη δημιουργία διανυσμάτων είναι η `rep`. Για παράδειγμα

```
> rep(v1, times=3)
 [1] 1 2 3 4 1 2 3 4 1 2 3 4
> rep(v1, each=3)
 [1] 1 1 1 2 2 2 3 3 3 4 4 4
> rep(v1, c(2,1,2,1))
[1] 1 1 2 3 3 4
```

Για να διατάξουμε τα δεδομένα ενός διανύσματος χρησιμοποιούμε τη συνάρτηση `sort`. Για να βρούμε πια συνιστώσα του αρχικού διανύσματος βρίσκεται σε κάθε συνιστώσα του διατεταγμένου διανύσματος χρησιμοποιούμε τη συνάρτηση `order`. Για παράδειγμα

```
> v4 <- c(1,10,123,64,32,55); sort(v4)
[1]   1  10  32  55  64 123
> order(v4)
[1] 1 2 5 6 4 3
```

Η *i* συνιστώσα ενός διανύσματος *x* επιλέγεται με το `x[i]`. Για παράδειγμα

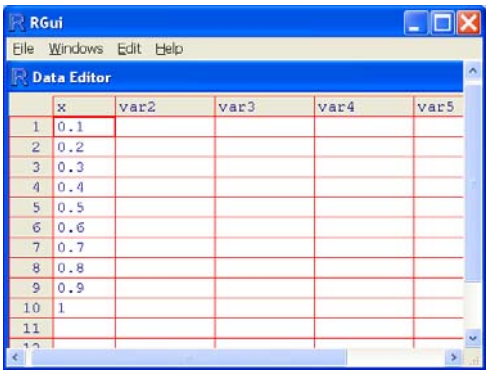
```
> x <- 1:10/10
> x[2]+2*x[3]          # 0.2+2*0.3
[1] 0.8
> x[2:5]
[1] 0.2 0.3 0.4 0.5
> x[c(4,6)]
[1] 0.4 0.6
> x[3] <- 11; x
[1]  0.1  0.2  11.0  0.4  0.5  0.6  0.7  0.8  0.9  1.0
```

Η αλλαγή της τιμής μιας ή περισσότερων συνιστωσών ενός διανύσματος μπορεί να επιτευχθεί και με χρήση "προγράμματος λογιστικού φύλου", με την εντολή

`data.entry(όνομα_διανύσματος)`.

Για παράδειγμα, η εκτέλεση της εντολής

`data.entry(x)`



ανοίγει ένα παράθυρο (Data Editor) όπου μπορούμε να προβούμε σε διορθώσεις των συνιστωσών του διανύσματος *x* κάνοντας διπλό κλικ στην συνιστώσα που θέλουμε να αλλάξουμε. Επίσης μπορεί να χρησιμοποιηθεί η διαδρομή Edit>Data editor..., ή η συνάρτηση `replace`. Η *i* συνιστώσα ενός διανύ-

σματος (ή λίστας) x διαγράφεται με $x[-i]$. Για παράδειγμα

```
> z <- seq(1,20,2);z[-4]
[1]  1  3  5  9 11 13 15 17 19
> y <- replace(z,3,55);y
[1]  1  3 55  7  9 11 13 15 17 19
```

Για να κρατήσουμε μόνο τα στοιχεία ενός διανύσματος x που ικανοποιούν μια συνθήκη γράφουμε `x[συνθήκη]`. Για παράδειγμα

```
> z[z<=4 | z>7]
[1]  1  3  9 11 13 15 17 19
```

Τα διανύσματα μπορούν να έχουν συνιστώσες και αλφαριθμητικές σειρές χαρακτήρων. Για παράδειγμα

```
> Οικογένεια <- c("Γεώργιος", "Άννα", "Λάζαρος", "Ελένη")
> Οικογένεια
[1] "Γεώργιος" "Άννα"     "Λάζαρος"  "Ελένη"
> Οικογένεια[2]
[1] "Άννα"
```

Μπορούμε να ενώσουμε διανύσματα με αριθμητικές και αλφαριθμητικές συνιστώσες. Το αποτέλεσμα είναι ένα διάνυσμα τύπου χαρακτήρα

```
> v5 <- c(Οικογένεια,v1); v5
[1] "Γεώργιος"  "Άννα"  "Λάζαρος"  "Ελένη"  "1"  "2"  "3"  "4"
> mode(v5)
[1] "character"
```

Στο σημείο αυτό αξίζει να αναφέρουμε ότι μπορούμε εύκολα να μετατρέψουμε μια σειρά δεδομένων σε χρονοσειρά με τη συνάρτηση `ts`. Για παράδειγμα

```
> x <- c(1:26)
> t1 <- ts(x, frequency = 4, start = c(1959, 3)); t1
     Qtr1 Qtr2 Qtr3 Qtr4
1959              1    2
1960    3    4    5    6
1961    7    8    9   10
1962   11   12   13   14
1963   15   16   17   18
1964   19   20   21   22
1965   23   24   25   26
> t2<- ts(x, frequency = 12, start = c(1959,6)); t2
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1959                         1   2   3   4   5   6   7
1960   8   9  10  11  12  13  14  15  16  17  18  19
1961  20  21  22  23  24  25  26
> mode(t2);class(t2)
[1] "numeric"
[1] "ts"
```

Κλείνοντας την παράγραφο που αφορά τα διανύσματα δίνουμε τον ακόλουθο πίνακα που περιέχει συναρτήσεις που εφαρμόζονται σε διανύσματα (εφαρμόζονται επίσης σε πίνακες, πλαίσια δεδομένων, κλπ.)

Πίνακας 2.1: Συναρτήσεις για σύνοψη δεδομένων

Συνάρτηση	Ερμηνεία
`sum(x)`	Άθροισμα στοιχείων
`cumsum(x)`	Προοδευτικό άθροισμα στοιχείων
`prod(x)`	Γινόμενο στοιχείων
`cumprod(x)`	Προοδευτικό γινόμενο στοιχείων
`max(x)`	Μέγιστο στοιχείο
`min(x)`	Ελάχιστο στοιχείο
`sort(x)`	Διάταξη των στοιχείων
`range(x)`	Διάνυσμα (min(x), max(x))
`length(x)`	Αριθμός στοιχείων
`mean(x)`	Μέσος
`median(x)`	Διάμεσος
`var(x)`	Διακύμανση
`sd(x)`	Τυπική απόκλιση
`skewness(x)`	Ασυμμετρία (πακέτο moments)
`kurtosis(x)`	Κύρτωση (πακέτο moments)
`cor(x,y)`	Συσχέτιση
`quantile(x)`	Ποσοστιαία σημεία
`rank(x)`	Βαθμοί
`IQR(x)`	Ενδοτεταρτημοριακό εύρος
`cov(x,y)`	Συνδιακύμανση
`summary(x)`	Προεπιλεγμένα μέτρα
`mode(x)`	Τύπος
`class(x)`	Τάξη

2.3 Πίνακες (matrices)

2.3.1 Βασικά στοιχεία

Ο πιο απλός τρόπος για να ορίσουμε ένα πίνακα στο R είναι μέσω της συνάρτησης `matrix`. Για παράδειγμα, η εισαγωγή του πίνακα $P = \begin{bmatrix} 2 & 1 & 7 \\ 4 & 5 & 3 \end{bmatrix}$ γίνεται ως εξής

```
> P <- matrix(c(2,4,1,5,7,3), nrow=2, ncol=3); P
     [,1] [,2] [,3]
[1,]    2    1    7
[2,]    4    5    3
> mode(P)
[1] "numeric"
> class(P)
[1] "matrix"
```

Στο ίδιο αποτέλεσμα θα καταλήγαμε (γιατί;) αν εκτελούσαμε κάθε μια από τις εντολές

```
> P <- matrix(c(2,4,1,5,7,3), nrow=2)
> P <- matrix(c(2,4,1,5,7,3), ncol=3)
```

Παρατηρούμε ότι τα στοιχεία του πίνακα δηλώνονται σε στήλες: πρώτα τα στοιχεία της πρώτης στήλης, μετά τα στοιχεία της δεύτερης στήλης, κ.ο.κ. Αν θέλουμε να δηλωθούν κατά γραμμές θα πρέπει να χρησιμοποιήσουμε το όρισμα `byrow=T`. Για παράδειγμα

```
> P <- matrix(c(2,1,7,4,5,3), nrow=2, ncol=3, byrow=T); P
     [,1] [,2] [,3]
[1,]    2    1    7
[2,]    4    5    3
```

Αν δοθούν λιγότερα στοιχεία σε ένα πίνακα από όσα απαιτούνται τότε τα στοιχεία που δεν ορίζονται συμπληρώνονται αυτόματα ξεκινώντας από την πρώτη δηλωμένη συνιστώσα. Για παράδειγμα

```
> M <- matrix(c(2,4,1,5), nrow=2, ncol=3)
Warning message:
In matrix(c(2, 4, 1, 5), nrow = 2, ncol = 3) :
  data length [4] is not a sub-multiple or multiple of the number of col-
umns [3]
> M
     [,1] [,2] [,3]
[1,]    2    1    2
[2,]    4    5    4
```

Για να δημιουργήσουμε ένα πίνακα με όλα τα στοιχεία του ίσα με κάποιο αριθμό *r* εκτελούμε την εντολή

```
> r <- 2
> matrix(r, nrow=2, ncol=3)
     [,1] [,2] [,3]
[1,]    2    2    2
[2,]    2    2    2
```

Οι πίνακες μπορούν να δημιουργηθούν από διανύσματα και με τη συνάρτηση `dim`. Για παράδειγμα

```
> va <- rep(1:4, each=3); va
 [1] 1 1 1 2 2 2 3 3 3 4 4 4
> dim(va) <- c(3,4); va
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    1    2    3    4
[3,]    1    2    3    4
```

Επίσης

```
> vb <- seq(-1.5, 1.5, length=12); vb
 [1] -1.5000000 -1.2272727 -0.9545455 -0.6818182 -0.4090909 -0.1363636
 [7]  0.1363636  0.4090909  0.6818182  0.9545455  1.2272727  1.5000000
> dim(vb) <- c(3,4); vb
           [,1]       [,2]      [,3]      [,4]
[1,] -1.5000000 -0.6818182 0.1363636 0.9545455
[2,] -1.2272727 -0.4090909 0.4090909 1.2272727
[3,] -0.9545455 -0.1363636 0.6818182 1.5000000
```

Μπορούμε να "ενώσουμε" πίνακες κατά γραμμές με τη συνάρτηση `rbind`, αλλά και κατά στήλες με τη συνάρτηση `cbind`, εφόσον το τελικό αποτέλεσμα είναι αποδεκτός πίνακας. Για παράδειγμα

```
> rb <- rbind(va,vb); rb
           [,1]       [,2]      [,3]      [,4]
[1,]  1.0000000  2.0000000 3.0000000 4.0000000
[2,]  1.0000000  2.0000000 3.0000000 4.0000000
[3,]  1.0000000  2.0000000 3.0000000 4.0000000
[4,] -1.5000000 -0.6818182 0.1363636 0.9545455
[5,] -1.2272727 -0.4090909 0.4090909 1.2272727
[6,] -0.9545455 -0.1363636 0.6818182 1.5000000
> cb <- cbind(va,vb); cb
     [,1] [,2] [,3] [,4]       [,5]       [,6]      [,7]      [,8]
[1,]    1    2    3    4 -1.5000000 -0.6818182 0.1363636 0.9545455
[2,]    1    2    3    4 -1.2272727 -0.4090909 0.4090909 1.2272727
[3,]    1    2    3    4 -0.9545455 -0.1363636 0.6818182 1.5000000
```

Για να δώσουμε ονόματα στις γραμμές και τις στήλες ενός πίνακα χρησιμοποιούμε τις συναρτήσεις `dimnames` και `list`. Για παράδειγμα

```
> Q <- matrix(1:20, nrow=4)
> dimnames(Q) <- list(c("R1","R2","R3","R4"),c("C1","C2","C3","C4","C5"))
> Q
   C1 C2 C3 C4 C5
R1  1  5  9 13 17
R2  2  6 10 14 18
R3  3  7 11 15 19
R4  4  8 12 16 20
```

Εναλλακτικά, χρησιμοποιούνται οι συναρτήσεις `rownames` και `colnames`. Για παράδειγμα

```
> Q <- matrix(1:20, nrow=4)
> rownames(Q) <- c("R1","R2","R3","R4")
> colnames(Q) <- c("C1","C2","C3","C4","C5")
> Q
   C1 C2 C3 C4 C5
R1  1  5  9 13 17
R2  2  6 10 14 18
R3  3  7 11 15 19
R4  4  8 12 16 20
```

Σε ένα πίνακα μπορούν να εφαρμοστούν οι συναρτήσεις `length`, `dim`, κλπ. Για παράδειγμα

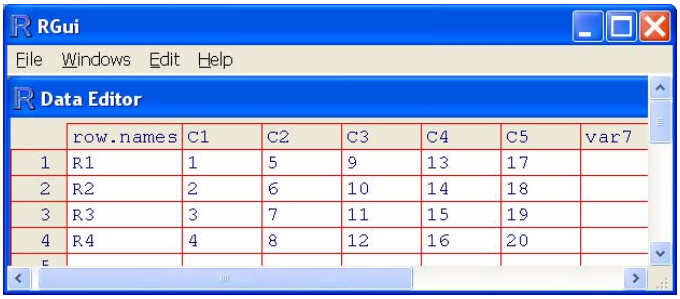
```
> dim(Q); length(Q); nrow(Q); ncol(Q)
[1] 4 5
[1] 20
[1] 4
[1] 5
> colMeans(Q)
[1]  2.5  6.5 10.5 14.5 18.5
> colSums(Q)
[1] 10 26 42 58 74
> rowMeans(Q)
[1]  9 10 11 12
> rowSums(Q)
[1] 45 50 55 60
```

Η επιλογή του (i, j) στοιχείου ενός πίνακα Q επιλέγεται με `Q[i,j]`, η επιλογή της i γραμμής γίνεται με `Q[i,]`, ενώ η επιλογή της j στήλης γίνεται με `Q[,j]`. Με το πρόσημο "-" γίνεται διαγραφή της αντίστοιχης γραμμής, στήλης. Για παράδειγμα

```
> Q[1,1]+Q[2,2]; Q[2,]; Q[,4]
[1] 7
[1]  2  6 10 14 18
[1] 13 14 15 16
> Q[-2,-4]
   C1 C2 C3 C5
R1  1  5  9 17
R3  3  7 11 19
R4  4  8 12 20
```

Η αλλαγή της τιμής μιας ή περισσότερων συνιστωσών ενός πίνακα μπορεί να επιτευχθεί με τη συνάρτηση `data.entry(όνομα_πίνακα)`, ή `edit(όνομα_πίνακα)`, ή με τη διαδρομή Edit>Data editor... . Η εκτέλεση της σχετικής εντολής ανοίγει τον Data Editor του R και έτσι μπορούμε να επέμβουμε στα στοιχεία του πίνακα και να τα αλλάξουμε αν επιθυμούμε. Για παράδειγμα

```
edit(Q)
```



2.3.2 Πράξεις με διανύσματα και πίνακες

Στον ακόλουθο πίνακα δίνονται σύμβολα πράξεων και συναρτήσεις που εφαρμόζονται σε διανύσματα ή/και πίνακες.

Πίνακας 2.2: Σύμβολα, συναρτήσεις που αφορούν διανύσματα και πίνακες

Σύμβολο - Συνάρτηση	Πράξη
`%*%`	Εσωτερικό γινόμενο διανυσμάτων Πολλαπλασιασμός πινάκων
`%o%`	Εξωτερικό γινόμενο διανυσμάτων
`t`	Ανάστροφος πίνακα
`solve`	Αντίστροφος πίνακα, λύση γραμμικού συστήματος
`diag`	Εξαγωγή της διαγωνίου ενός πίνακα Κατασκευή διαγώνιου πίνακα με διαγώνιο ένα διάνυσμα Κατασκευή μοναδιαίου πίνακα
`eigen`	Ιδιοτιμές και ιδιοδιανύσματα πίνακα

Για διανύσματα δίνουμε το κάτωθι παράδειγμα

```
> u <- c(1,2,3); v <- c(2,3,4); u; v
[1] 1 2 3
[1] 2 3 4
> u+v; u*v          # πρόσθεση και πολλαπλασιασμός στοιχείο προς στοιχείο
[1] 3 5 7
[1] 2 6 12
> cbind(u)             # το διάνυσμα u σαν πίνακας στήλη
     u
[1,] 1
[2,] 2
[3,] 3
> rbind(u)             # το διάνυσμα u σαν πίνακας γραμμή
  [,1] [,2] [,3]
u    1    2    3
> u%*%v                # εσωτερικό γινόμενο
     [,1]
[1,]   20
> rbind(u)%*%cbind(v)  # ξανά εσωτερικό γινόμενο
   v
u 20
> rbind(u)%*%rbind(v)  # μη επιτρεπτός πολλαπλασιασμός πινάκων
Error in rbind(u) %*% rbind(v) : non-conformable arguments
> u%o%v                # εξωτερικό γινόμενο
     [,1] [,2] [,3]
[1,]    2    3    4
[2,]    4    6    8
[3,]    6    9   12
> cbind(u)%*%rbind(v)  # ξανά εξωτερικό γινόμενο (επίσης outer(u,v,"*"))
     [,1] [,2] [,3]
[1,]    2    3    4
[2,]    4    6    8
[3,]    6    9   12
```

Αξίζει να σημειώσουμε ότι η πράξη $x\%*\%x$ μπορεί να σημαίνει είτε $x'x$ είτε xx' όπου το x είναι ένα διάνυσμα στήλη. Το R όμως δίνει σαν αποτέλεσμα τον πίνακα με τη μικρότερη διάσταση, δηλαδή δίνει ως αποτέλεσμα το εσωτερικό γινόμενο. Για παράδειγμα

```
> x <- c(1,1,2)
> x%*%x
     [,1]
[1,]    6
```

Για πίνακες δίνουμε το κάτωθι παράδειγμα

```
> A <- matrix(c(2,3,4,5,7,6,2,4,1), nrow=3, ncol=3)
> B <- matrix(c(1,5,7,3,1,2,5,7,8), nrow=3)
> w <- c(1,2,3)
> A
     [,1] [,2] [,3]
[1,]    2    5    2
[2,]    3    7    4
[3,]    4    6    1
```

```
> B
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    5    1    7
[3,]    7    2    8
> w
[1] 1 2 3
> A*A                  # πολλαπλασιασμός στοιχείο προς στοιχείο (A²[i,j])
     [,1] [,2] [,3]
[1,]    4   25    4
[2,]    9   49   16
[3,]   16   36    1
> w%*%A
     [,1] [,2] [,3]
[1,]   20   37   13
> A%*%w
     [,1]
[1,]   18
[2,]   29
[3,]   19
> A%*%B
     [,1] [,2] [,3]
[1,]   41   15   61
[2,]   66   24   96
[3,]   41   20   70
> t(B)                  # ανάστροφος του πίνακα B
     [,1] [,2] [,3]
[1,]    1    5    7
[2,]    3    1    2
[3,]    5    7    8
> solve(A)              # αντίστροφος του πίνακα A
           [,1]       [,2]       [,3]
[1,] -1.5454545  0.6363636  0.5454545
[2,]  1.1818182 -0.5454545 -0.1818182
[3,] -0.9090909  0.7272727 -0.0909091
> diag(A)               # η διαγώνιος του πίνακα A
[1] 2 7 1
> diag(3)               # μοναδιαίος πίνακας διαστάσεων 3X3
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> diag(w)               # διαγώνιος πίνακας με διαγώνιο το διάνυσμα w
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    2    0
[3,]    0    0    3
```

Για τον υπολογισμό των ιδιοτιμών και των ιδιοδιανυσμάτων[7] ενός πίνακα δίνουμε το ακόλουθο παράδειγμα

[7] Μια ιδιοτιμή λ και το αντίστοιχο ιδιαδιάνυσμα x ενός πίνακα A ικανοποιούν την εξίσωση $Ax=\lambda x$. Οι ιδιοτιμές προκύπτουν ως λύσεις της εξίσωσης $|A-\lambda I|=0$.

```
> C <- matrix(c(2,-4,2,8), nrow=2); C
     [,1] [,2]
[1,]    2    2
[2,]   -4    8
> OBJ <- eigen(C); OBJ
$values
[1] 6 4

$vectors
           [,1]       [,2]
[1,] -0.4472136 -0.7071068
[2,] -0.8944272 -0.7071068
```

Εκχωρήσαμε στο αντικείμενο `OBJ` το αποτέλεσμα της συνάρτησης `eigen` που είναι δύο στοιχεία: οι ιδιοτιμές (`$values`) και τα ιδιοδιανύσματα (`$vectors`) που δίνονται ως στήλες ενός πίνακα. Για να καλέσουμε οποιοδήποτε από τα δύο στοιχεία γράφουμε το όνομα του αντικειμένου (`OBJ`) και το όνομα του στοιχείου σαν μια λέξη. Για να εμφανίσουμε τη λίστα των "ιδιοτήτων" ενός αντικειμένου χρησιμοποιούμε τη συνάρτηση `attributes`. Για παράδειγμα

```
> sum(OBJ$val)
[1] 10
> diag(OBJ$vec)
[1] -0.4472136 -0.7071068
> attributes(OBJ)
$names
[1] "values"  "vectors"
> C%*%OBJ$vectors[,2]
          [,1]
[1,] -2.828427
[2,] -2.828427
> OBJ$values[2]*OBJ$vectors[,2]
[1] -2.828427 -2.828427
```

Για να επιλύσουμε το γραμμικό σύστημα $Ax = b$ χρησιμοποιούμε τη συνάρτηση `solve(A,b)` που επιστρέφει το διάνυσμα $x = A^{-1}b$. Για την επίλυση του συστήματος

$$\begin{aligned} 2x &+ 3y &+ 4z &= 20 \\ &\quad 4y &+ 3z &= 17 \\ x &+ 2y & &= 5 \end{aligned}$$

που είναι το διάνυσμα $(x, y, z) = (1,2,3)$, έχουμε

```
> A <- matrix(c(2,0,1,3,4,2,4,3,0), nrow=3)
> b <- c(20,17,5)
> solve(A,b)
[1] 1 2 3
```

2.4 Πίνακες μεγαλύτερης διάστασης (arrays)

Τα διανύσματα έχουν μία διάσταση και οι πίνακες δύο διαστάσεις. Στο R συλλογές αριθμών με $k \geq 3$ διαστάσεις δημιουργούνται με τη συνάρτηση `array` (τα διανύσματα και οι πίνακες είναι ar-

rays). Για παράδειγμα

```
> ar <- array(c(1:10,21:30,41:44), dim=c(2,3,4)); ar
, , 1

     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

, , 2

     [,1] [,2] [,3]
[1,]    7    9   21
[2,]    8   10   22

, , 3

     [,1] [,2] [,3]
[1,]   23   25   27
[2,]   24   26   28

, , 4

     [,1] [,2] [,3]
[1,]   29   41   43
[2,]   30   42   44
```

Το αντικείμενο `ar` είναι ένα array διαστάσεων $2 \times 3 \times 4$. Τα στοιχεία ενός array μπορούν να επιλεχθούν όπως τα στοιχεία των διανυσμάτων και των πινάκων. Για παράδειγμα

```
> ar[,,1]
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> ar[2,,1]
[1] 2 4 6
> ar[,2,1]
[1] 3 4
> ar[1,,]
     [,1] [,2] [,3] [,4]
[1,]    1    7   23   29
[2,]    3    9   25   41
[3,]    5   21   27   43
> ar[,2,]
     [,1] [,2] [,3] [,4]
[1,]    3    9   25   41
[2,]    4   10   26   42
```

Στα array εφαρμόζονται οι γνωστές συναρτήσεις που χρησιμοποιούνται και στους πίνακες. Για παράδειγμα

```
> length(ar); dim(ar); mode(ar); class(ar); nrow(ar); ncol(ar)
[1] 24
[1] 2 3 4
[1] "numeric"
[1] "array"
```

```
[1] 2
[1] 3
> dimnames(ar) <- list(c("R1","R2"),c("C1","C2","C3"),
+ c("ar1","ar2","ar3","ar4"))
> ar
, , ar1

   C1 C2 C3
R1  1  3  5
R2  2  4  6

, , ar2

   C1 C2 C3
R1  7  9 21
R2  8 10 22

, , ar3

   C1 C2 C3
R1 23 25 27
R2 24 26 28

, , ar4

   C1 C2 C3
R1 29 41 43
R2 30 42 44
```

2.5 Παράγοντες (factors)

Μια ποιοτική (ή κατηγορική) μεταβλητή δηλώνεται στο R με τη συνάρτηση `factor`. Για παράδειγμα, έστω ότι δίνονται οι τιμές του λογάριθμου του αριθμού των λευκών αιμοσφαιρίων σε 20 ασθενής (μεταβλητή $\log wbc$). Αν $\log wbc \le 2.1$, $2.1 < \log wbc \le 3$, $\log wbc > 3$ το επίπεδο (μεταβλητή *lmh*) του αριθμού των λευκών αιμοσφαιρίων χαρακτηρίζεται ως χαμηλό (*Lo*), μέσο (*Med*), υψηλό (*Hi*), αντίστοιχα. Η μεταβλητή *lmh* είναι κατηγορική (παράγοντας, *factor*) με τρία επίπεδα (*Lo*, *Med*, *Hi*) και υπάρχει ιεράρχηση στα επίπεδά της (*levels*). Τα δεδομένα μας δίνονται στο ακόλουθου πλαίσιο

$\log wbc$	2.3	1.4	0.6	2.1	2.2	2.4	2.3	2.6	1.0	2.7
lmh	Med	Lo	Lo	Lo	Med	Med	Med	Med	Lo	Med
$\log wbc$	2.2	3.1	2.8	3.7	0.8	3.8	3.9	1.1	2.6	1.7
lmh	Med	Hi	Med	Hi	Lo	Hi	Hi	Lo	Med	Lo

Οι δύο παραπάνω μεταβλητές εισάγονται στο R ως εξής

```
> logwbc <- c(2.3,1.4,0.6,2.1,2.2,2.4,2.3,2.6,1.0,2.7,
+ 2.2,3.1,2.8,3.7,0.8,3.8,3.9,1.1,2.6,1.7)
> v <- c("Med","Lo","Lo","Lo","Med","Med","Med","Med","Lo","Med",
+ "Med","Hi","Med","Hi","Lo","Hi","Hi","Lo","Med","Lo")
```

```
> v
 [1] "Med" "Lo"  "Lo"  "Lo"  "Med" "Med" "Med" "Med" "Lo"  "Med"
[11] "Med" "Hi"  "Med" "Hi"  "Lo"  "Hi"  "Hi"  "Lo"  "Med" "Lo"
> mode(v)
[1] "character"
> lmh <- factor(v)            # μετατροπή του διανύσματος v σε παράγοντα
> lmh
 [1] Med Lo  Lo  Lo  Med Med Med Med Lo  Med Med Hi  Med Hi  Lo
[16] Hi  Hi  Lo  Med Lo
Levels: Hi Lo Med
> mode(lmh)
> class(lmh)
[1] "numeric"
[1] "factor"
```

Ένας πιο κομψός τρόπος εισαγωγής του διανύσματος *v* (με το όνομα *x*) είναι ο ακόλουθος

```
> x <- c(1:20)
> x[logwbc<=2.1]<-"Lo"
> x[logwbc>2.1&logwbc<=3]<-"Med"
> x[logwbc>3]<-"Hi"
> x
 [1] "Med" "Lo"  "Lo"  "Lo"  "Med" "Med" "Med" "Med" "Lo"  "Med"
[11] "Med" "Hi"  "Med" "Hi"  "Lo"  "Hi"  "Hi"  "Lo"  "Med" "Lo"
```

Με τη συνάρτηση `tapply` μπορούμε να εφαρμόσουμε ειδικές συναρτήσεις σε κάθε ομάδα στοιχείων του πρώτου ορίσματος που δηλώνονται από τα επίπεδα του δεύτερου ορίσματος. Για παράδειγμα

```
> logwbcmeans <- tapply(logwbc, lmh, mean)
> logwbcmeans
      Hi        Lo       Med
3.625000  1.242857  2.455556
> logwbcsum <- tapply(logwbc,lmh,sum)
> logwbcsum
  Hi   Lo  Med
14.5  8.7 22.1
```

Η συνάρτηση `ordered` είναι ανάλογη της `factor`, μόνο που με την πρώτη μπορούμε να ιεραρχήσουμε τα επίπεδα ενός παράγοντα. Για παράδειγμα

```
> lmh <- ordered(v, levels=c("Lo","Med","Hi"))
> lmh
 [1] Med Lo  Lo  Lo  Med Med Med Med Lo  Med Med Hi  Med Hi  Lo  Hi  Hi
Lo  Med Lo
Levels: Lo < Med < Hi
```

2.6 Λίστες (lists)

Όλα τα αντικείμενα δεδομένων που γνωρίσαμε μέχρι τώρα ήταν του ιδίου τύπου. Η λίστα (list) είναι μια διατεταγμένη (και αριθμημένη) συλλογή αντικειμένων που μπορεί να είναι διαφορετικών τύπων. Για τη δημιουργία μιας λίστας δεδομένων χρησιμοποιείτε η εντολή `list`. Για παράδειγμα

```
> a <- "Δημήτρης"; b <- "Μαρία"; c <- 3
>list(a,b,c)
[[1]]
[1] "Δημήτρης"

[[2]]
[1] "Μαρία"

[[3]]
[1] 3
```

Δημιουργήσαμε μια λίστα με τρία στοιχεία. Ονόματα στα στοιχεία της λίστας δίνονται ως εξής

```
> LstA <- list(father.name=a, mother.name=b, no.children=c)
> LstA
$father.name
[1] "Δημήτρης"

$mother.name
[1] "Μαρία"

$no.children
[1] 3

> mode(LstA)
[1] "list"
```

Για να ενώσουμε δύο λίστες χρησιμοποιείται η συνάρτηση c. Για παράδειγμα

```
> LstB <- list(child.names=c("Παναγιώτης", "Ελένη", "Λάζαρος"),
+ child.ages=c(5, 11, 12))
> LstB
$child.names
[1] "Παναγιώτης"  "Ελένη"  "Λάζαρος"

$child.ages
[1]  5 11 12

> Lst <- c(LstA,LstB)
> Lst
$father.name
[1] "Δημήτρης"

$mother.name
[1] "Μαρία"

$no.children
[1] 3

$child.names
[1] "Παναγιώτης"  "Ελένη"  "Λάζαρος"

$child.ages
[1]  5 11 12

> attributes(Lst)
$names
```

```
[1] "father.name" "mother.name" "no.children" "child.names" "child.ages"

> names(Lst)
[1] "father.name" "mother.name" "no.children" "child.names" "child.ages"
```

Η επιλογή του *i* κατά σειρά στοιχείου μιας λίστας *L* επιλέγεται με `L[[i]]` ή με `L$odject_name`. Για παράδειγμα

```
> length(Lst)
[1] 5
> Lst[[5]]
[1]  5 11 12
> Lst$child.ages
[1]  5 11 12
> Lst[[5]][1:2]
[1]  5 11
> Lst[5]              # επιλογή του 5ου στοιχείου μαζί με το όνομά του
$child.ages
[1]  5 11 12

> sum(Lst[[5]])
[1] 28
> sum(Lst$child.ag)
[1] 28
> sum(Lst[5])
Error in sum(Lst[5]) : invalid 'type' (list) of argument
```

Εφαρμογή συναρτήσεων στα στοιχεία μια λίστας γίνεται με τη συνάρτηση `lapply`.

```
> lapply(LstA,length)
$father.name
[1] 1

$mother.name
[1] 1

$no.children
[1] 1

> lapply(LstB,mean)
$child.names
[1] NA

$child.ages
[1] 9.333333
```

Η αλλαγή της τιμής μιας ή περισσότερων συνιστωσών μιας λίστας μπορεί να γίνει με την εντολή `data.entry(όνομα_λίστας)`.

2.7 Πλαίσια δεδομένων (data frames)

Το πλαίσιο δεδομένων είναι μια λίστα τάξης "data.frame". Η βασική διαφορά με τις λίστες είναι ότι όλα τα στοιχεία ενός πλαισίου δεδομένων έχουν το ίδιο μήκος. Τα πλαίσια δεδομένων εισάγονται

με τη συνάρτηση `data.frame`. Για παράδειγμα

```
> a <- c("male","female","male","female","male","male")
> b <- c(24,32,45,67,43,21)
> c <- c(181,167,178,170,175,2008)
> d <- c(81,55,75,74,78,95)
> df <- data.frame(SEX=a,AGE=b,HEIGHT=c,WEIGHT=d)
> df
     SEX AGE HEIGHT WEIGHT
1   male  24    181     81
2 female  32    167     55
3   male  45    178     75
4 female  67    170     74
5   male  43    175     78
6   male  21   2008     95
> mode(df); class(df)
[1] "list"
[1] "data.frame"
> names(df)
[1] "SEX"    "AGE"    "HEIGHT" "WEIGHT"
> mode(df$SEX); class(df$SEX)
[1] "numeric"
[1] "factor"
> dim(df); length(df)
[1] 6 4
[1] 4
```

Για να αλλάξουμε τα ονόματα των μεταβλητών σε ένα πλαίσιο δεδομένων χρησιμοποιούμε τη συνάρτηση `names`.

```
> names(df) <- c("Col1","Col2","Col3","Col4")
> df
    Col1 Col2 Col3 Col4
1   male   24  181   81
2 female   32  167   55
3   male   45  178   75
4 female   67  170   74
5   male   43  175   78
6   male   21 2008   95
```

Για να προσθέσουμε μια στήλη σε ένα πλαίσιο δεδομένων χρησιμοποιούμε τη συνάρτηση `cbind`, ενώ για να προσθέσουμε μια γραμμή στο τέλος του πλαισίου δεδομένων χρησιμοποιούμε τη συνάρτηση `rbind`. Για παράδειγμα

```
> e <- c("Y","N","N","N","N","Y")
> DF <- cbind(df,SMOKE=e)
> DF
     SEX AGE HEIGHT WEIGHT SMOKE
1   male  24    181     81     Y
2 female  32    167     55     N
3   male  45    178     75     N
4 female  67    170     74     N
5   male  43    175     78     N
6   male  21   2008     95     Y
```

Σε αρκετές περιπτώσεις η επιλογή μιας μεταβλητής (στήλης) ενός πλαισίου δεδομένων με χρήση του συμβόλου "$", για παράδειγμα `df$SEX` ή `DF$SMOKE`, δεν είναι πάντοτε βολική. Θα θέλαμε να χρησιμοποιούμε τις μεταβλητές με το όνομά τους. Αυτό επιτυγχάνεται με τη συνάρτηση `attach`. Για παράδειγμα

```
> sum(WEIGHT)
Error: object "WEIGHT" not found
> attach(df)
> sum(WEIGHT)
[1] 458
```

Για να διαπιστώσουμε ποια data.frames έχουν επισυναφθεί (attached) εκτελούμε την εντολή `search` (δείτε επίσης Παράγραφο 1.7). Επίσης για να "αποσυνάψουμε" ένα data.frame χρησιμοποιούμε τη συνάρτηση `detach`.

```
> search()
 [1] ".GlobalEnv"        "df"                "package:stats"
 [4] "package:graphics"  "package:grDevices" "package:utils"
 [7] "package:datasets"  "package:methods"   "Autoloads"
[10] "package:base"
```

Για να διατάξουμε ένα πλαίσιο δεδομένων ως προς μια στήλη του χρησιμοποιούμε τη συνάρτηση `order`. Για παράδειγμα

```
> df[order(AGE,decreasing=TRUE),]
     SEX AGE HEIGHT WEIGHT
4 female  67    170     74
3   male  45    178     75
5   male  43    175     78
2 female  32    167     55
1   male  24    181     81
6   male  21   2008     95
> df[order(SEX,AGE),]
     SEX AGE HEIGHT WEIGHT
2 female  32    167     55
4 female  67    170     74
6   male  21   2008     95
1   male  24    181     81
5   male  43    175     78
3   male  45    178     75
```

Για την ένωση δύο πλαισίων δεδομένων χρησιμοποιείται η συνάρτηση `merge`. Για παράδειγμα

```
> a1 <- c("male","female")
> b1 <- c(23,33)
> c1 <- c(182, 163)
> d1 <- c(57, 71)
> e1 <- c("Y", "N")
> df1 <- data.frame(SEX=a1,AGE=b1,HEIGHT=c1,WEIGHT=d1, SMOKE=e1)
> merge(df, df1, all=T)
     SEX AGE HEIGHT WEIGHT SMOKE
1 female  32    167     55  <NA>
2 female  33    163     71     N
3 female  67    170     74  <NA>
```

```
4   male  21   2008     95  <NA>
5   male  23    182     57     Y
6   male  24    181     81  <NA>
7   male  43    175     78  <NA>
8   male  45    178     75  <NA>
```

Για την επιλογή στοιχείων σε ένα πλαίσιο δεδομένων δίνουμε το ακόλουθο παράδειγμα

```
> df[2,2]
[1] 32
> df[1,]
   SEX AGE HEIGHT WEIGHT
1 male  24    181     81
> df[,"HEIGHT"]
[1]  181  167  178  170  175 2008
> df[SEX=="male",]
   SEX AGE HEIGHT WEIGHT
1 male  24    181     81
3 male  45    178     75
5 male  43    175     78
6 male  21   2008     95
```

Το ύψος του έκτου ατόμου με τιμή 2008 είναι προφανώς λανθασμένη καταχώρηση. Ας υποθέσουμε ότι το πραγματικό ύψος είναι 192. Η διόρθωση τη τιμής του ύψους γίνεται ως εξής

```
> HEIGHT[6] <- 192
> HEIGHT
[1] 181 167 178 170 175 192
> df
     SEX AGE HEIGHT WEIGHT
1   male  24    181     81
2 female  32    167     55
3   male  45    178     75
4 female  67    170     74
5   male  43    175     78
6   male  21   2008     95
```

Παρατηρούμε ότι η μεταβλητή HEIGHT έχει διορθωθεί όχι όμως και η αντίστοιχη καταχώρηση στο πλαίσιο δεδομένων. Για να γίνει η αλλαγή και στο πλαίσιο δεδομένων εκτελούμε τις εντολές

```
> df$HEIGHT[6] <- 192
> df
     SEX AGE HEIGHT WEIGHT
1   male  24    181     81
2 female  32    167     55
3   male  45    178     75
4 female  67    170     74
5   male  43    175     78
6   male  21    192     95
```

Εναλλακτικά, για να διορθώσουμε ένα πλαίσιο δεδομένων θα μπορούσαμε να χρησιμοποιήσουμε τη συνάρτηση `edit(όνομα_πλαισίου_δεδομένων)` που ανοίγει τον Data Editor του R.

	SEX	AGE	HEIGHT	WEIGHT	var5	var
1	male	24	181	81		
2	female	32	167	55		
3	male	45	178	75		
4	female	67	170	74		
5	male	43	175	78		
6	male	21	192	95		
7						

2.8 Σύνοψη εντολών Κεφαλαίου 2

```
array, assign, attach, attributes
c, cbind, class, colnames, cor, cos, cov, cumprod, cumsum
data.entry, data.frame, detach, diag, dim, dimnames
edit, eigen
factor
IQR
lapply, length, list
matrix, max, mean, median, merge, min, mode
names, ncol, nrow
order, ordered, outer
prod
quantile
rank, range, rbind, rep, replace, rev, rownames
scan, seq, sd, solve, sort, sum, summary
tapply, ts
var
```

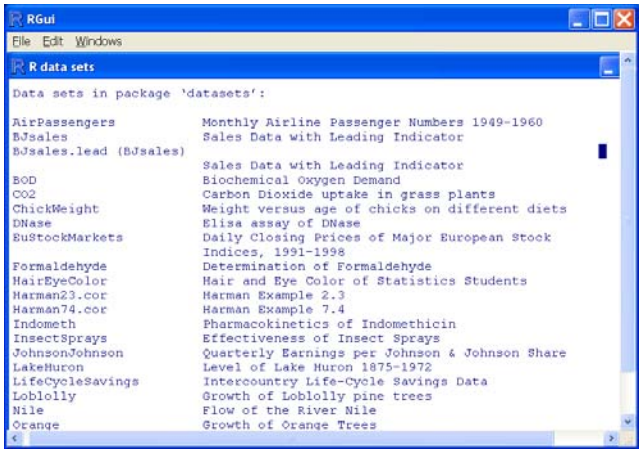
ΚΕΦΑΛΑΙΟ 3

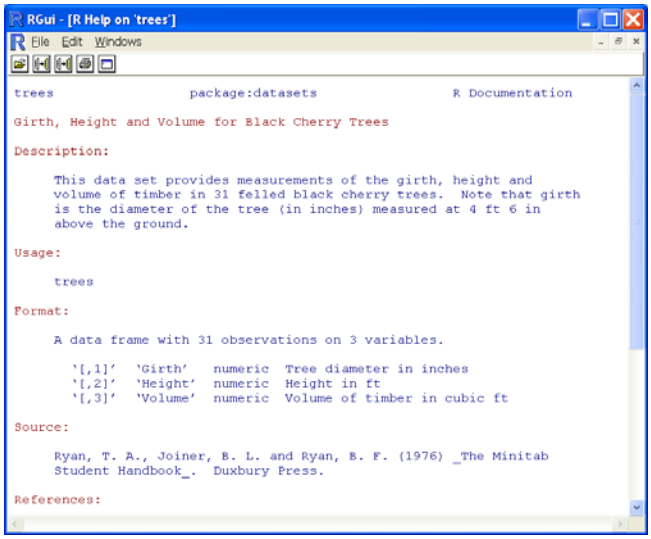
Γραφήματα

3.1 Συνάρτηση plot

Οι συναρτήσεις `demo(graphics)`, `demo(persp)` και `demo(image)` παρέχουν μια επίδειξη των γραφικών δυνατοτήτων του R. Η πιο συνηθισμένη συνάρτηση για σχεδίαση στο R είναι η συνάρτηση `plot`. Για να επιδείξουμε τη συνάρτηση θα ανοίξουμε το αρχείο δεδομένων `trees` του R. Μια περιγραφή του αρχείου και των μεταβλητών του προκύπτει με τη συνάρτηση `help(trees)` ή `?trees`. Τα διαθέσιμα σύνολα δεδομένων του R εμφανίζονται με τη συνάρτηση `data`.

```
> data()
> help(trees)
```



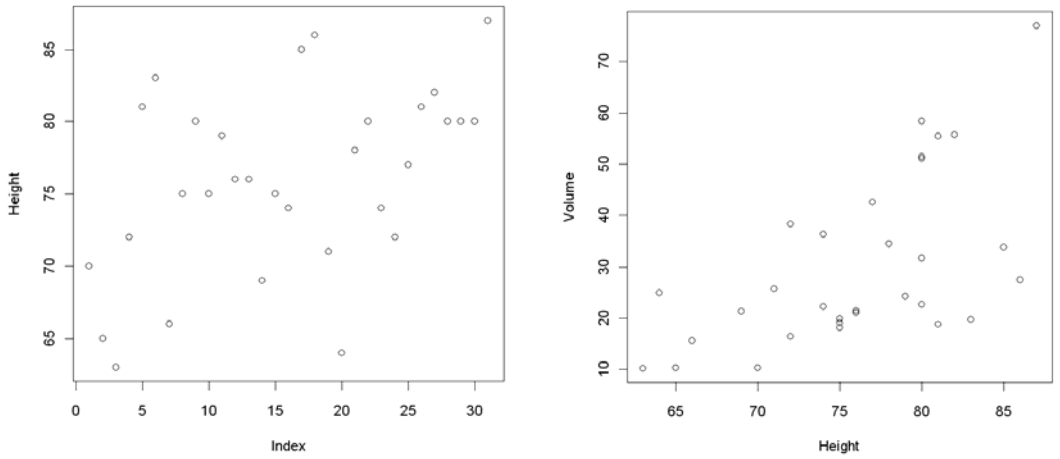


Ακολούθως δίνεται η ταυτότητα του συνόλου δεδομένων `trees`.

```
> trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
          .
          .
          .
30   18.0     80   51.0
31   20.6     87   77.0
> mode(trees); class(trees)
[1] "list"
[1] "data.frame"
```

Εκτελώντας τις ακόλουθες εντολές παίρνουμε ένα index γράφημα της μεταβλητής Height και ένα διάγραμμα διασποράς των μεταβλητών (Height, Volume).

```
> attach(trees)
> plot(Height)
> plot(Height,Volume)
```



Η συνάρτηση `plot` δέχεται αρκετά ορίσματα που καθορίζουν την ακριβή εμφάνιση του γραφήματος. Τα ορίσματα αυτά περιγράφονται στη βοήθεια της συνάρτησης `plot` και της συνάρτησης `par`. Ο ακόλουθος πίνακας περιλαμβάνει τα σημαντικότερα από αυτά.

Πίνακας 3.1: Τα βασικότερα ορίσματα της συνάρτησης plot

`main="κείμενο"`	Τίτλος γραφήματος
`sub="κείμενο"`	Υπότιτλος γραφήματος
`xlab="κείμενο"`	Τίτλος του άξονα των x
`ylab="κείμενο"`	Τίτλος του άξονα των y
`xlim=c(a,b)`	Όρια του άξονα των x (διάστημα $[a,b]$)
`ylim=c(a,b)`	Όρια του άξονα των y (διάστημα $[a,b]$)
`cex=u`	Το u είναι διάνυσμα (ή ένας αριθμός) που δηλώνει το σχετικό μέγεθος των σημείων και του κειμένου του γραφήματος
`cex.axis, cex.lab`	Σχετικό μέγεθος γραμματοσειράς αξόνων, τίτλων αξόνων
`cex.main, cex.sub`	Σχετικό μέγεθος γραμματοσειράς τίτλου, υπότιτλου
`font=n`	Καθορίζει τον τύπο της γραμματοσειράς του γραφήματος ($n = 1,2,...,5$, π.χ. $2 = \text{bold},\ 3 = \text{italic}$)
`font.axis, font.lab`	Τύπος γραμματοσειράς αξόνων, τίτλων αξόνων
`font.main, font.sub`	Τύπος γραμματοσειράς τίτλου, υπότιτλου
`tck=a`	Μήκος των κάθετων σημαδιών (tick marks) στους άξονες (συνήθως $-0.05 < a < 0.05$, $default = -0.02$)

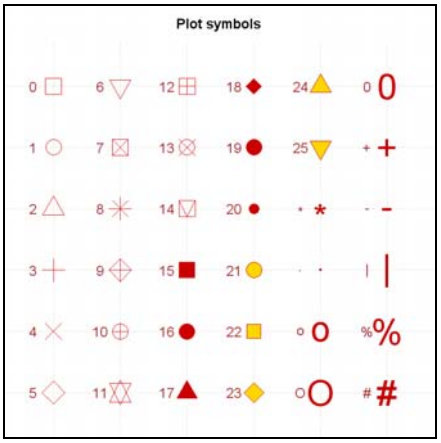
`lab=c(n1,n2,n3)`	$n1$ $(n2)$: Αριθμός κάθετων σημαδιών στον x (y) άξονα $n3$: Αριθμός χαρακτήρων για τα κάθετα σημάδια (μη εφαρμόσιμο)
`las=n`	Προσανατολισμός των ετικετών (label) των αξόνων ($n = 0,1,2$)
`xaxt=n`	Ο άξονας των x δεν σχεδιάζεται
`yaxt=n`	Ο άξονας των y δεν σχεδιάζεται
`frame.plot=FALSE`	Δεν σχεδιάζεται το πλαίσιο στην περιοχή γραφήματος
`lty="n"`	Τύπος γραμμής ($n = 1,2,...,6$)
`lwd="a"`	Πάχος γραμμής $(a > 0)$
`asp=n`	Aspect ratio y / x
`log="w"`	Άξονας w σε λογαριθμική κλίμακα ($w = x, y$)
`log="xy"`	Άξονας των x , y σε λογαριθμική κλίμακα
`pch="c"`	Σύμβολο χαρακτήρας στη θέση των σημείων του γραφήματος ($c = a,b,c,\ldots,A,B,C,\ldots,1,2,\ldots$)
`pch=n` ή `pch="σύμβολο"`	Σύμβολο στη θέση των σημείων "○" του γραφήματος ($n = 1,2,3,...$ π.χ. $10 = \otimes$, $18 = \blacklozenge$)
`col="χρώμα"`	Χρώμα γραμμής ή/και σημείων του γραφήματος (Τα 657 διαθέσιμα χρώματα δίνονται με τη συνάρτηση `colors()`)
`col=n`	Χρώμα γραμμής ή/και σημείων του γραφήματος ($n = 1,2,..,8$, π.χ. $2 =$ κόκκινο, $7 =$ κίτρινο)
`col.axis, col.lab`	Χρώμα αξόνων, τίτλων αξόνων, τίτλου, υπότιτλου
`col.main, col.sub`	Χρώμα, τίτλου, υπότιτλου
`bg="χρώμα"` ή `bg=n`	Χρώμα του φόντου, ή του φόντου του `legeng` ή εσωτερικό χρώμα των σημείων του γραφήματος (με `pch=21:25`

Οι δυνατές τιμές του ορίσματος `type` που ορίζει τον τύπο της γραμμής σε ένα γράφημα δίνονται στον ακόλουθο πίνακα

Πίνακας 3.2: Το όρισμα type

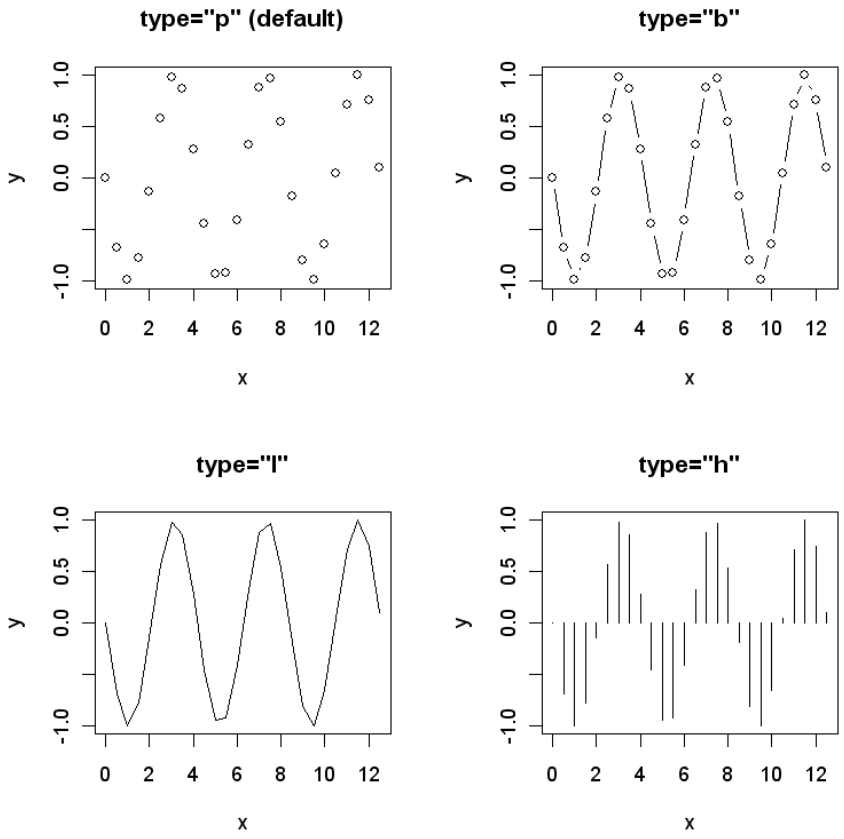
`type="p"`	Σημεία (default)
`type="l"`	Γραμμή
`type="b"`	Γραμμή και σημεία (μη επικαλυπτόμενα)
`type="c"`	Γραμμή με κενό στα σημεία
`type="o"`	Γραμμή και σημεία ενωμένα (επικαλυπτόμενα)
`type="h"`	Κάθετες γραμμές
`type="s"`	Σκαλοπάτι (οριζόντιο και μετά κάθετο)
`type="S"`	Σκαλοπάτι (κάθετο και μετά οριζόντιο)
`type="n"`	Τίποτα

Στο ακόλουθο πλαίσιο δίνονται τα σύμβολα που μπορούν να χρησιμοποιηθούν ως σύμβολα απεικόνισης των σημείων ενός γραφήματος με χρήση του ορίσματος `pch`.



Για να επιδείξουμε τη χρήση του Πίνακα 3.2 δίνουμε το ακόλουθο παράδειγμα

```
> x <- seq(0,4*pi,0.5)
> y <- -sin(1.5*x)
> par(mfcol=c(2,2))
> plot(x,y,main="type=''p'' (default)")
> plot(x,y,type="l", main="type=''b''")
> plot(x,y,type="b", main="type=''l''")
> plot(x,y,type="h", main="type=''h''")
> par(mfcol=c(1,1))
```

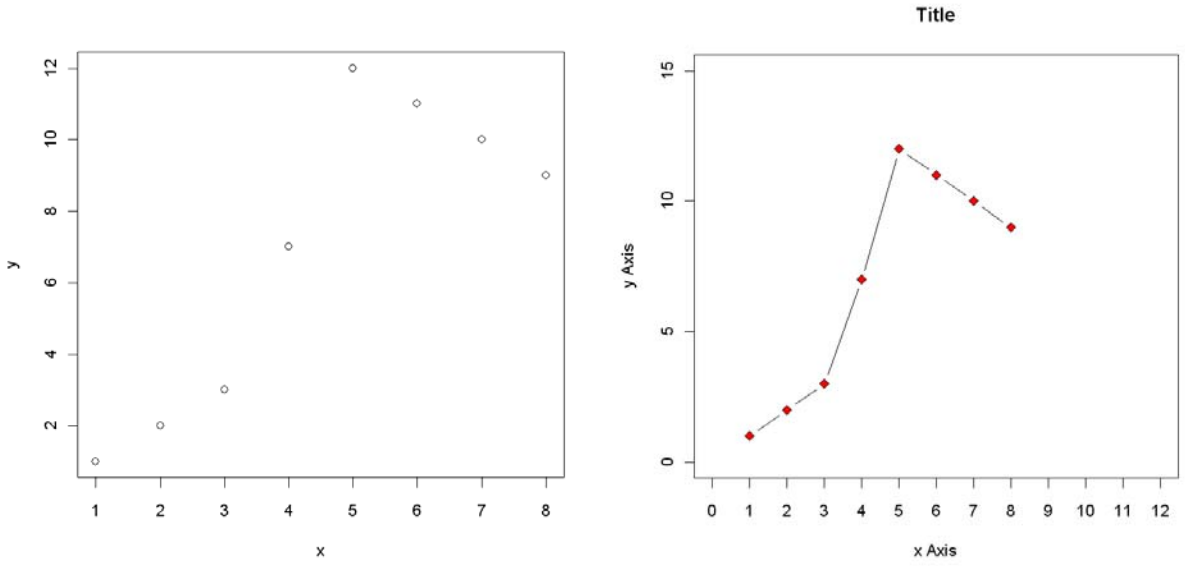


Η εκτέλεση της εντολής > `par(mfcol=c(2,2))` δημιουργεί ένα πάνελ διαστάσεων 2×2 στο οποίο εισάγονται τα τέσσερα γραφήματα διαδοχικά σε στήλες. Στο τέλος είναι απαραίτητη η εκτέλεση της εντολής > `par(mfcol=c(1,1))` προκειμένου το επόμενο γράφημα να σχεδιαστεί σε όλο το διαθέσιμο χώρο του παράθυρου των γραφικών. Σημειώνουμε ότι η δήλωση παραμέτρων του γραφήματος μέσω της συνάρτησης `par` είναι μόνιμη. Για να επαναφέρουμε τις αρχικές ρυθμίσεις εκτελούμε τις κάτωθι εντολές

```
> oldpar <- par(no.readonly=TRUE); par(oldpar)
```

Για την επίδειξη ορισμένων ορισμάτων του Πίνακα 3.1 δίνουμε το ακόλουθο παράδειγμα

```
> x <- c(1:8)
> y <- c(1:3, 7, 12:9)
> plot(x,y)
> plot(x,y, type="b", pch=23, bg="red", main="Title", xlab="x Axis",
+ ylab="y Axis", xlim=c(0,12), ylim=c(0,15), lab=c(13,4,7))
```



3.2 Συναρτήσεις χαμηλού επιπέδου

Η συνάρτηση `plot` θεωρείται συνάρτηση γραφημάτων υψηλού επιπέδου (high level) και κάθε φορά που χρησιμοποιείται δημιουργεί συνήθως ένα νέο γράφημα (το παλιό διαγράφεται). Αρκετές φορές σε ένα γράφημα θέλουμε να ενσωματώσουμε διάφορα άλλα στοιχεία χωρίς όμως αυτό να διαγραφεί. Αυτό επιτυγχάνεται με τις αποκαλούμενες συναρτήσεις γραφημάτων χαμηλού επιπέδου (low level). Μερικές από αυτές δίνονται στον ακόλουθο πίνακα

Πίνακας 3.3: Βασικές συναρτήσεις χαμηλού επιπέδου

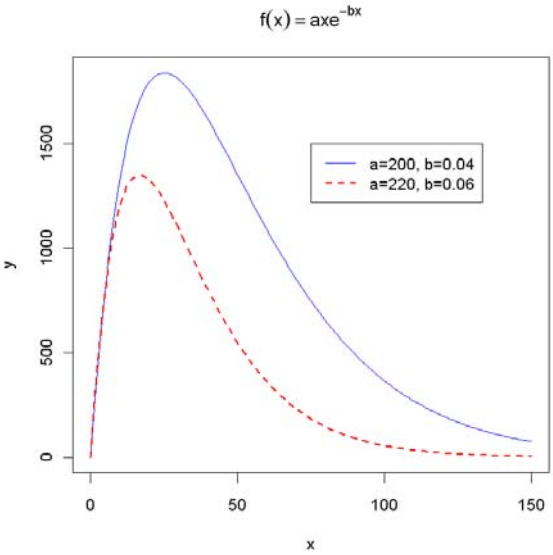
`points(x,y)`	Εισάγει νέα σημεία στο γράφημα
`lines(x,y)`	Εισάγει καμπύλες στο γράφημα
`text(x,y,label="abc")`	Εισάγει το κείμενο το `"abc"` στη θέση (x, y)

`segments(x0,y0,x1,y1)`	Ενώνει με γραμμή τα σημεία $(x0, y0)$, $(x1, y1)$
`abline(a,b)`	Εισάγει τη γραμμή $y = a + bx$
`title("abc")`	Εισάγει το κείμενο το `"abc"` ως τίτλο του γραφήματος
`rug(x)`	Εισάγει μικρές κάθετες γραμμές πάνω στον άξονα των x που αντιστοιχούν στα σημεία του γραφήματος
`rect(x0,y0,x1,y1)`	Εισάγει ένα ορθογώνιο με διαγώνιο που ορίζεται από τα σημεία $(x0, y0)$, $(x1, y1)$
`arrows(x0,y0,x1,y1)`	Εισάγει ένα τόξο που ορίζεται από τα σημεία $(x0, y0)$, $(x1, y1)$
`legend(x,y,...)`	Εισάγει υπόμνημα στη θέση (x, y)

Για παράδειγμα

```
> x <- seq(0,150,0.5)
> y1 <- 200*x*exp(-0.04*x)
> y2 <- 220*x*exp(-0.06*x)
> plot(c(x,x),c(y1,y2), type="n", xlab="x",ylab="y")      #1
> lines(x,y1,lty=1,col="blue")                            #2
> lines(x,y2,lty=2,col="red", lwd=2)                      #3
> title(expression(f(x)==a*x*e^{-b*x}))                   #4
> legend(75, 1500, c("a=200, b=0.04", "a=220, b=0.06"), lty=c(1,2),
+ col=c("blue","red"), lwd=c(1,2))                        #5
```

Με την εντολή #1 δημιουργείται η περιοχή του γραφήματος χωρίς να εμφανιστούν οι καμπύλες (`type="n"`), ενώ ταυτόχρονα δίνονται τίτλοι στους άξονες. Με την εντολή #2 προστίθεται στην περιοχή του γραφήματος η γραφική παράσταση της συνάρτησης $f(x) = 200xe^{-0.04x}$ με γραμμή χρώματος μπλε (`col="blue"`), και με την εντολή #3 προστίθεται η γραφική παράσταση της συνάρτησης $f(x) = 220xe^{-0.06x}$ με διακεκομμένη γραμμή (`lty=2`) χρώματος κόκκινου (`col="red"`) πάχους 2 (`lwd=2`). Με την εντολή #4 προστίθεται ο τίτλος του γραφήματος. Αντί της εντολής `title("f(x)=a*x*exp(b*x)")` που θα έβαζε ως τίτλο την ακριβή έκφραση που περικλείεται στα `" "`, προτιμήθηκε η μαθηματική γραφή[8] μέσω της συνάρτησης `expression`. Με την εντολή #5 προστίθεται υπόμνημα στη θέση



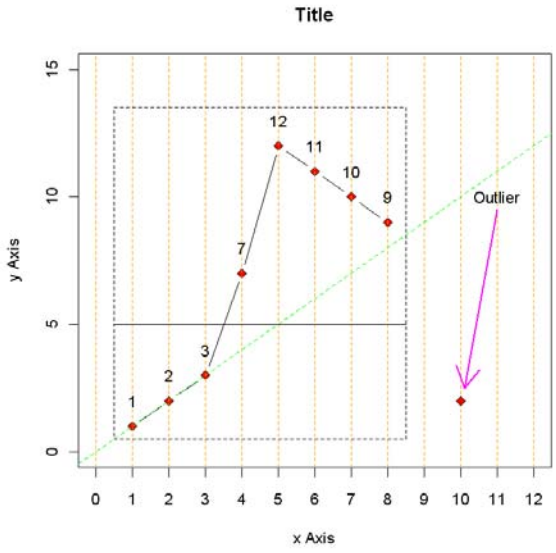
[8] Περισσότερες πληροφορίες για τη γραφή μαθηματικών εκφράσεων σε γραφήματα παρέχουν οι συναρτήσεις `help(plotmath)`, `example(plotmath)` και `demo(plotmath)`.

με συντεταγμένες $(75,1500)$.

Επίσης με συναρτήσεις γραφημάτων χαμηλού επιπέδου μπορούμε να εισάγουμε σημεία, γραμμές, πλέγμα, κείμενο και πολύγωνα στο γράφημά μας. Για παράδειγμα

```
> x <- c(1:8)
> y <- c(1:3, 7, 12:9)
> plot(x,y)
> plot(x,y, type="b", pch=23, bg="red", main="Title", xlab="x Axis",
+ ylab="y Axis", xlim=c(0,12), ylim=c(0,15), lab=c(13,4,7))
grid(nx = NULL, ny = NA, col = "orange", lty = 2)                #1
> text(x,y,labels=y, pos=3, offset=1)                            #2
> points(10,2, pch=23, bg="red")                                 #3
> text(11,10,label="Outlier")                                    #4
> arrows(11,9.5,10.1,2.5,col=6, lwd=2)                           #5
> abline(0,1, lty=2, col="green")                                #6
> rect(0.5,0.5,8.5,13.5, lty=2)                                  #7
> segments(0.5,5,8.5,5)                                          #8
```

Με την εντολή #1 δημιουργείται πλέγμα στην περιοχή του γραφήματος μόνο για τον άξονα x (`nx=NULL, ny=NA`). Με την εντολή #2 δίνονται τιμές στα σημεία (x,y) του γραφήματος και μάλιστα οι τιμές που αντιστοιχούν στην τεταγμένη y (`labels=y`). Οι τιμές δίνονται στο πάνω μέρος των σημείων (`pos=3`) και απέχουν απόσταση ενός χαρακτήρα (`offset=1`). Με την εντολή #3 προστίθεται ένα νέο σημείο, της ίδιας μορφής με τα υπόλοιπα σημεία του γραφήματος, στη θέση $(10,2)$. Με την εντολή #4 εισάγεται το κείμενο Outlier στη θέση $(11,10)$. Με την εντολή #5 άγεται ένα διάνυσμα από το σημείο $(11,9.5)$ στο σημείο $(x,y)=(10.1,2.5)$. Με την εντολή #6, προστίθεται μια γραμμή στο γράφημα που τέμνει τον άξονα των y στο 0 και έχει κλίση ίση με 1 (η γραμμή που ορίζει η συνάρτηση `abline(a,b)` είναι η $y=a+bx$). Με την εντολή #7 εισάγεται στο γράφημα ένα ορθογώνιο που ορίζεται δίνοντας δύο σημεία που καθορίζουν μια διαγώνιό του. Με την εντολή #8 εισάγεται στο γράφημα ένα ευθύγραμμο τμήμα που ορίζεται δίνοντας τα δύο σημεία που καθορίζουν τα άκρα του.

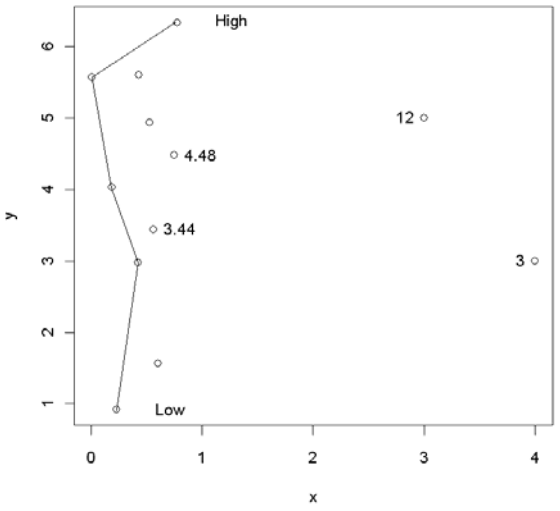


Το R προσφέρει διαδραστικές δυνατότητες επεξεργασίας γραφημάτων. Για παράδειγμα

```
> x <- c(0.6, 0.01, 4, 0.56, 0.43, 0.53, 0.78, 0.19, 0.23, 0.42, 0.75, 3)
> y <- c(1.56, 5.57, 3, 3.44, 5.60, 4.94, 6.34, 4.03, 0.92, 2.97, 4.48,
5)
> plot(x,y)
> identify(x,y)                          #1
> text(locator(2), c("High", "Low"))     #2
```

```
> identify(x,y,labels=y)                    #3
> locator( ,type="l")                      #4
```

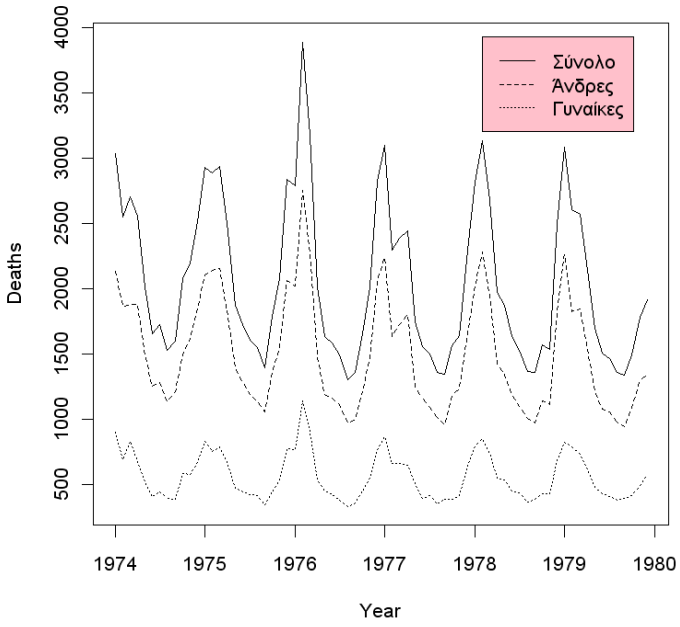
Όταν εκτελεστεί η εντολή #1 τότε ο δρομέας παίρνει τη μορφή "+" όταν βρεθεί πάνω από την περιοχή γραφήματος. Κάνοντας αριστερό κλικ πάνω στα σημεία "○" του γραφήματος εμφανίζεται ο αύξων αριθμός των σημείων (στην περίπτωσή μας το $12^{ο}$ (3,5) και το $3^{ο}$ (4,3)). Για να τερματιστεί η διαδικασία κάνουμε δεξί κλικ στην περιοχή του γραφήματος και επιλέγουμε Stop. Με την εντολή #2 εισάγεται κείμενο σε δύο σημεία της περιοχής του γραφήματος που θα επιλέξουμε. Η συνάρτηση `locator` φέρει το ίδιο αποτέλεσμα με την εντολή `identify(x,y)`, δηλαδή μετατρέπει το δρομέα σε "+" επιτρέποντας την επιλογή σημείων της περιοχής του γραφήματος. Με την εντολή #3 εισάγονται οι τεταγμένες των σημείων "○" του γραφήματος που θα επιλέξουμε. Με την εντολή #4 ενώνουμε σημεία της περιοχής του γραφήματος με μια τεθλασμένη γραμμή (εδώ επιλέξαμε 5 σημεία "○" του γραφήματος).



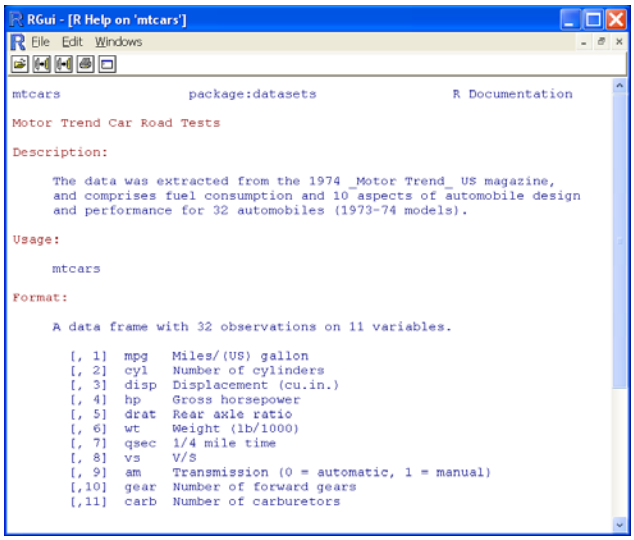
3.3 Συναρτήσεις ts.plot, pairs, matplot

Για να σχεδιάσουμε στο ίδιο γράφημα αρκετές χρονοσειρές που έχουν το ίδιο μέγεθος χρησιμοποιούμε τη συνάρτηση `ts.plot`. Χαρακτηριστικό παράδειγμα στο R είναι οι χρονοσειρές (σύνολα δεδομένων τάξης `ts`) `mdeaths`, `fdeaths` και `ldeaths` που δίνουν των αριθμό των μηνιαίων θανάτων των ανδρών, των γυναικών και στο σύνολο αντίστοιχα, λόγω βρογχίτιδας, εμφυσήματος και άσθματος στη Μεγάλη Βρετανία από το 1974 έως το 1979. Για παράδειγμα

```
> class(ldeaths); class(mdeaths); class(fdeaths)
[1] "ts"
[1] "ts"
[1] "ts"
> ts.plot(ldeaths,mdeaths,fdeaths,lty=1:3, xlab="Year", ylab="Deaths")
> leg.names <- c("Σύνολο","Άνδρες","Γυναίκες")
> legend(locator(1), leg.names, lty=1:3, bg="pink")
```

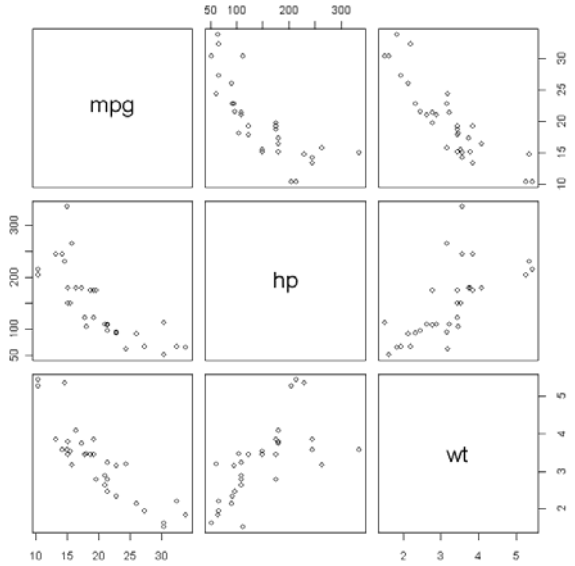


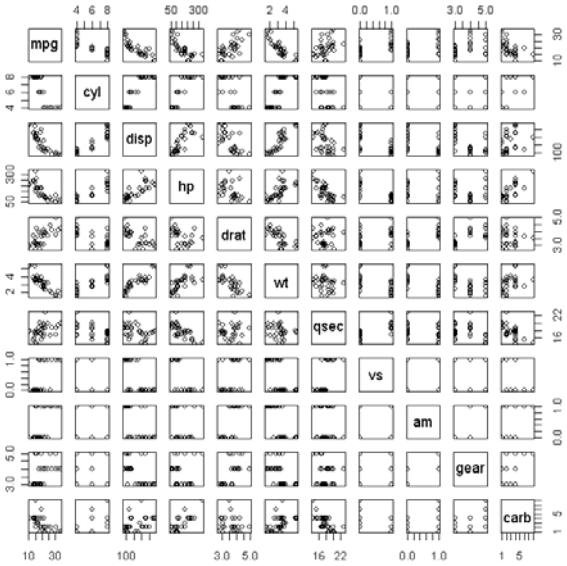
Το αρχείο δεδομένων `mtcars` του R περιέχει καταναλώσεις καυσίμων και άλλα 10 χαρακτηριστικά για 32 μοντέλα αυτοκινήτων (κατασκευής 1973-1974). Για να κατασκευαστούν διαγράμματα διασπορών για κάθε ζευγάρι μεταβλητών ενός συνόλου μεταβλητών χρησιμοποιείται η συνάρτηση `pairs`. Για παράδειγμα, το διάγραμμα διασποράς των μεταβλητών `mpg` (μίλια ανά γαλόνι), `hp` (ιπποδύναμη) και `wt` (βάρος σε λίβρες/1000) προκύπτει ως ακολούθως




```
> ?mtcars
> class(mtcars)
[1] "data.frame"
> attach(mtcars)
> pairs(cbind(mpg,hp,wt))
```

Για να πάρουμε το ολοκληρωμένο διάγραμμα διασποράς, επειδή το αρχείο δεδομένων `mtcars` είναι τάξης data.frame, αρκεί να εκτελεστεί η εντολή `plot(mtcars)`.

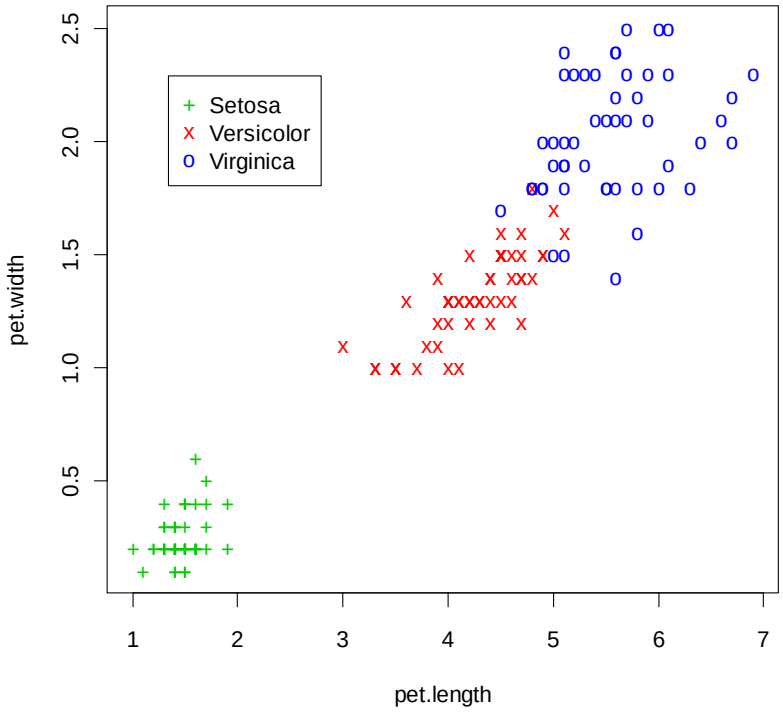




Για τη σχεδίαση των στηλών ενός πίνακα έναντι των στηλών ενός άλλου χρησιμοποιείται η η συνάρτηση `matplot`. Για παράδειγμα θα χρησιμοποιήσουμε το αρχείο δεδομένων `iris3` του R, που είναι τάξης array και δίνει μετρήσεις σε εκατοστά των μεταβλητών μήκος και πλάτος του σεπάλου και του πετάλου 50 λουλουδιών τριών ειδών κρίνων (*setosa*, *versicolor* και *virginica*). Εκτελώντας τις ακόλουθες εντολές προκύπτει διάγραμμα διασποράς του μήκους του πετάλου έναντι του πλάτους του πετάλου για τα τρία είδη των κρίνων

```
> iris3
, , Setosa
      Sepal L. Sepal W. Petal L. Petal W.
 [1,]      5.1      3.5      1.4      0.2
 [2,]      4.9      3.0      1.4      0.2
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
[50,]      5.0      3.3      1.4      0.2
, , Versicolor
      Sepal L. Sepal W. Petal L. Petal W.
 [1,]      7.0      3.2      4.7      1.4
 [2,]      6.4      3.2      4.5      1.5
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
 [50,]      5.7      2.8      4.1      1.3
, , Virginica
      Sepal L. Sepal W. Petal L. Petal W.
 [1,]      6.3      3.3      6.0      2.5
 [2,]      5.8      2.7      5.1      1.9
 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
 [50,]      5.9      3.0      5.1      1.8
```

```
> class(iris3)
[1] "array"
> pet.length <- iris3[,3,]  # επιλογή τρίτης στήλης από παντού
> class(pet.length)
[1] "matrix"
> pet.width <- iris3[,4,]   # επιλογή τέταρτης στήλης από παντού
> class(pet.width)
[1] "matrix"
> matplot(pet.length, pet.width, pch = c("+","x","o"), col=c(3,2,4))
> leg.names <- c("Setosa","Versicolor","Virginica")
> legend(locator(1), leg.names, pch = c("+","x","o"), col=c(3,2,4))
```



3.4 Σύνοψη εντολών Κεφαλαίου 3

```
abline, arrows
colors
data, demo(graphics),demo(persp) demo(image)
expression
identify
legend, lines, locator
matplot
pairs, par, plot, points
rect, rug
segments
text, title, ts.plot.
```

ΚΕΦΑΛΑΙΟ 4

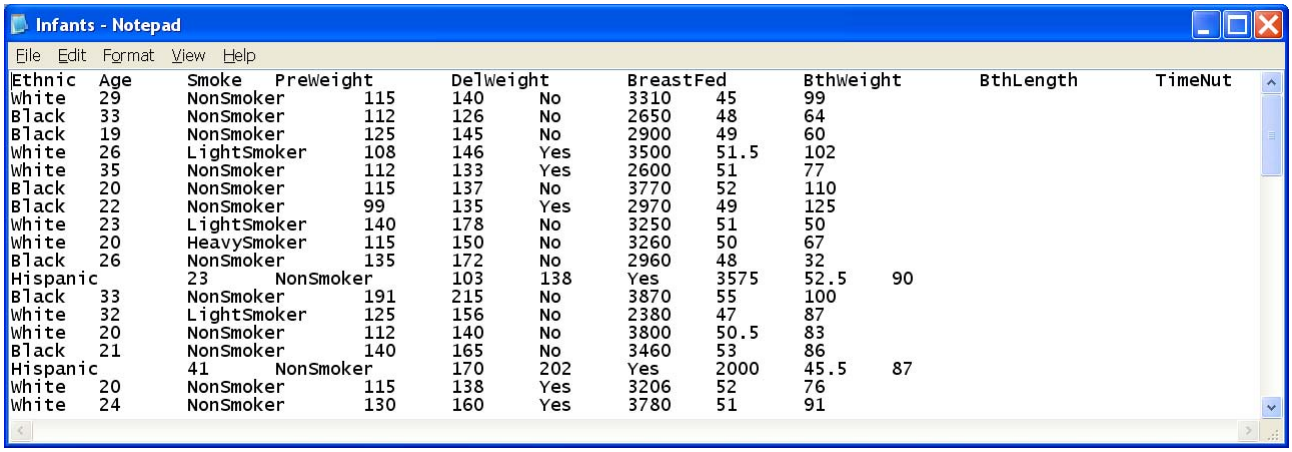
Στοιχεία Περιγραφικής Στατιστικής

4.1 Εισαγωγή δεδομένων από εξωτερικά αρχεία

Τα περισσότερα στατιστικά πακέτα προσφέρουν τη δυνατότητα αποθήκευσης αρχείων δεδομένων σε μορφή txt. Η εισαγωγή δεδομένων από αρχεία τύπου txt στο R γίνεται με τη συνάρτηση `read.table`. Για παράδειγμα, ας θεωρήσουμε αρχείο Infants.txt που βρίσκεται στον κατάλογο

C:/Documents and Settings/DLA/Desktop

το οποίο περιέχει δεδομένα που σχετίζονται με την απόφαση μιας μητέρας να θηλάσει το βρέφος της (68 γυναίκες, 9 μεταβλητές).



Η εισαγωγή των δεδομένων του αρχείου Infants.txt στο R γίνεται ως ακολούθως

```
> a <-read.table("C:/Documents and Settings/DLA/Desktop/infants.txt",
+ header=TRUE)
> a
    Ethnic Age       Smoke PreWeight DelWeight BreastFed BthWeight BthLength TimeNut
1    White  29   NonSmoker       115       140        No      3310      45.0      99
2    Black  33   NonSmoker       112       126        No      2650      48.0      64

.....................................................................................

68   Black  19   NonSmoker       132       156        No      3360      51.0      84
> class(a)
[1] "data.frame"
> names(a)
[1] "Ethnic"    "Age"       "Smoke"     "PreWeight" "DelWeight" "Breast-
Fed" "BthWeight" "BthLength" "TimeNut"
```

Ωστόσο, αν το αρχείο Infants.txt βρίσκεται στο working directory (δείτε Παράγραφο .6) τότε αρκεί η εκτέλεση της εντολής

```
a <- read.table("infants.txt", header=TRUE)
```

Τα δεδομένα του αρχείου Infants.txt μεταφέρονται στο αντικείμενο a που είναι ένα πλαίσιο δεδομένων (data.frame).

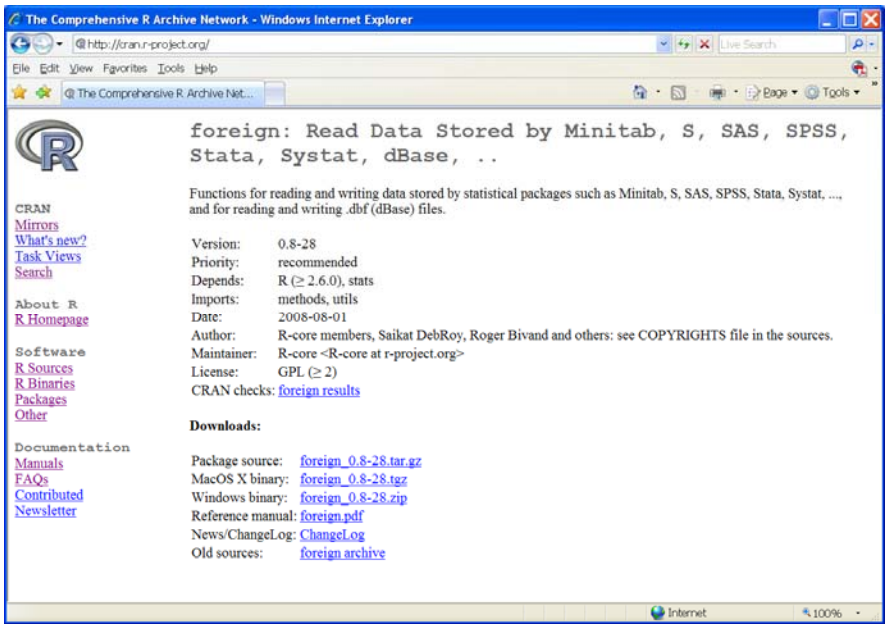
Η εισαγωγή αρχείων δεδομένων από στατιστικά πακέτα (SPSS, MINITAB, S-Plus, κτλ.), χωρίς πρώτα να μετατραπούν σε κάποια άλλη μορφή αρχείων, γίνεται με τη βοήθεια του πακέτου foreign. Για παράδειγμα, το αρχείο Cars.sav του στατιστικού πακέτου SPSS βρίσκεται στον κατάλογο

C:\Program Files\SPSSInc\PASWStatistics18\Samples\English

Η εισαγωγή των δεδομένων του αρχείου στο R γίνεται με τη συνάρτηση `read.spss`.

```
> library(foreign)
b <-read.spss("C:/ProgramFiles/SPSSInc/PASWStatistics18/Samples/English
+ /Cars.sav", use.value.labels=TRUE, to.data.frame = TRUE)
b
     MPG ENGINE HORSE WEIGHT ACCEL        YEAR   ORIGIN    CYLINDER     FILTER_.
1   18.0  307.0   130   3504  12.0          70 American 8 Cylinders Not Selected
2   15.0  350.0   165   3693  11.5          70 American 8 Cylinders Not Selected

.....................................................................................

406 31.0  119.0    82   2720  19.4          82 American 4 Cylinders     Selected
> class(b)
[1] "data.frame"
```

Για τη χρήση αλλά και για τις δυνατότητες που προσφέρει το πακέτο foreign ο ενδιαφερόμενος αναγνώστης παραπέμπεται στο εγχειρίδιο χρήσης του πακέτου (επισκεφτείτε την ιστοσελίδα http://cran.r-project.org/, επιλέξτε Packages, Table of available packages, sorted by name και κατόπιν foreign και foreign.pdf).



Μπορούμε να φορτώσουμε δεδομένα που περιέχονται σε προεγκατεστημένα πακέτα του R, χωρίς να φορτωθούν τα πακέτα, με τη συνάρτηση `data`. Για παράδειγμα

```
> data(HUMMER, package="UsingR")   # Φόρτωση του αρχείου δεδομένων HUMMER
από το πακέτο UsingR χωρίς να φορτωθεί το πακέτο
> HUMMER
```

```
      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
2003                          2493 2654 2987 2837 3157 2837 3157
2004 1927 2141 2334 2268 1982 2175 2505 2404 2548 2554 2693 3814
2005 1864 1866 2220 1700 2964 6754 7476 6367 5806 5640 5991 8079
2006 5214 5645
```

Η εκτέλεση της εντολής `library(UsingR)` φορτώνει το πακέτο μαζί με τα δεδομένα που υπάρχουν σε αυτό.

Κλείνοντας την παρούσα παράγραφο αξίζει να αναφέρουμε ότι μπορούμε να σώσουμε πλαίσια δεδομένων ή πίνακες σε μορφήASCII με τη συνάρτηση με τη συνάρτηση `write.table` (δείτε επίσης τις συναρτήσεις dump και `source`). Για παράδειγμα, μπορούμε να αποθηκεύσουμε τα δεδομένα του πλαισίου δεδομένων b ως εξής

```
write.table(b, file="C:/Documents and Settings/DLA/Desktop/car.txt")
```

4.2 Ανάλυση δεδομένων: Μια μεταβλητή

Η μεταβλητή Smoke του αρχείου Infants.txt δηλώνει αν μια μητέρα είναι μη καπνίστρια (NonSmoker), αν καπνίζει αρκετά (HeavySmoker) ή αν καπνίζει λίγο (LightSmoker).

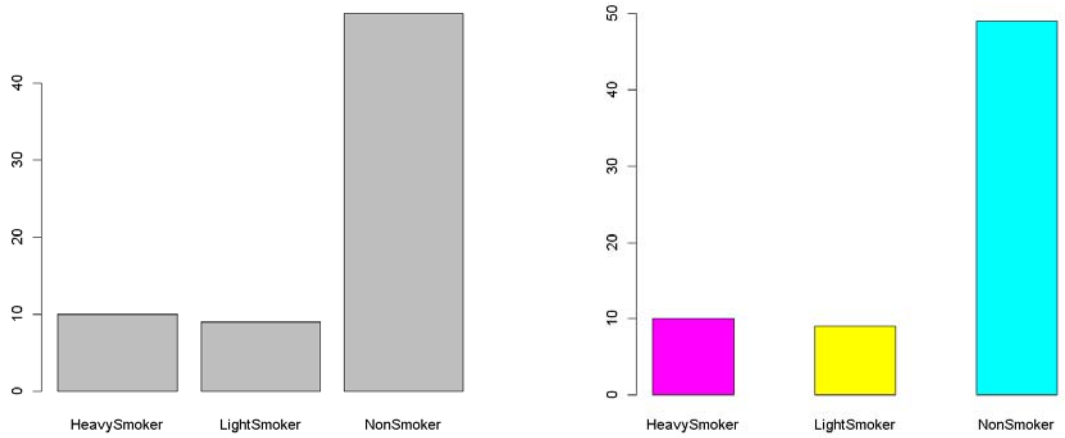
```
> attach(a)
> class(Smoke)
[1] "factor"
> length(Smoke)
[1] 68
> Smoke
 [1] NonSmoker   NonSmoker   NonSmoker   LightSmoker NonSmoker   NonSmoker
 [7] NonSmoker   LightSmoker HeavySmoker NonSmoker   NonSmoker   NonSmoker
.....................................................................
 [67] NonSmoker   NonSmoker
Levels: HeavySmoker LightSmoker NonSmoker
```

Η μεταβλητή Smoke είναι κατηγορική μεταβλητή (categorical). Τέτοιου είδους μεταβλητές παρουσιάζονται σε πίνακες ενώ συνηθίζεται και η γραφική παράστασή τους με ραβδογράμματα (barcharts) και διαγράμματα πίτας (pie charts). Οι συχνότητες κάθε τιμής της κατηγορικής μεταβλητής Smoke βρίσκονται με τη συνάρτηση `table`. Για παράδειγμα

```
> table(Smoke)
Smoke
HeavySmoker LightSmoker   NonSmoker
         10           9          49
```

Η κατασκευή ραβδογράμματος επιτυγχάνεται με τη συνάρτηση `barplot`. Για παράδειγμα δίνουμε τα ακόλουθα δύο ραβδογράμματα

```
> barplot(table(Smoke))
> barplot(table(Smoke), ylim=c(0,50), col=c(6,7,5), space=1)
```

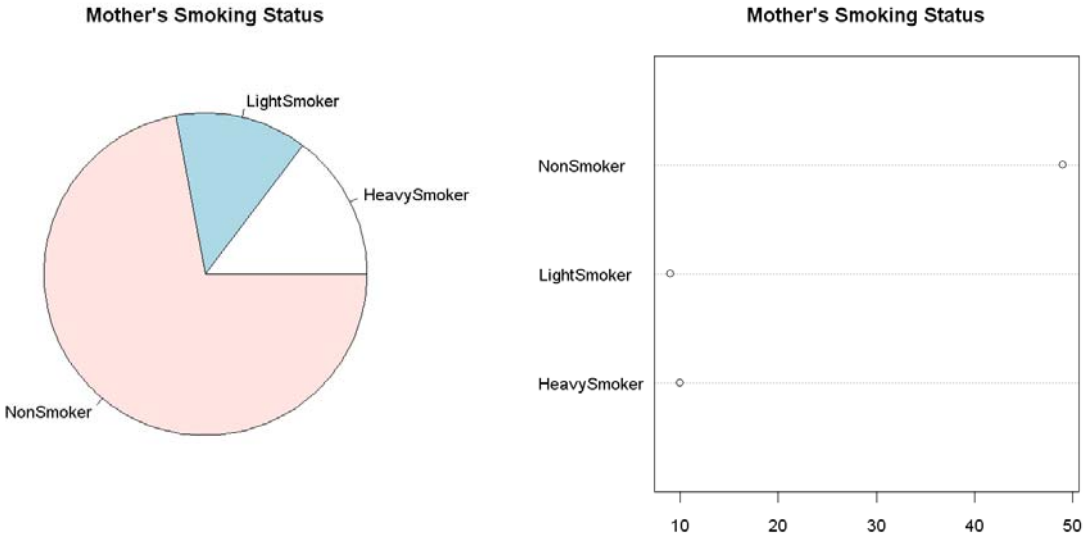


Το τελευταίο ραβδόγραμμα θα μπορούσε να προκύψει (με περισσότερο κόπο) ως εξής

```
Smoke.counts <- c(10,9,49)
names(Smoke.counts) <- c("HeavySmoker", "LightSmoker", "NonSmoker")
barplot(Smoke.counts, ylim=c(0,50), col=c(6,7,5))
```

Με τις συναρτήσεις `pie` και `dotchart` κατασκευάζεται διάγραμμα πίτας και διάγραμμα κουκκίδων, αντίστοιχα.

```
> pie(Smoke.counts, main="Mother's Smoking Status")
> dotchart(Smoke.counts, main="Mother's Smoking Status")
```



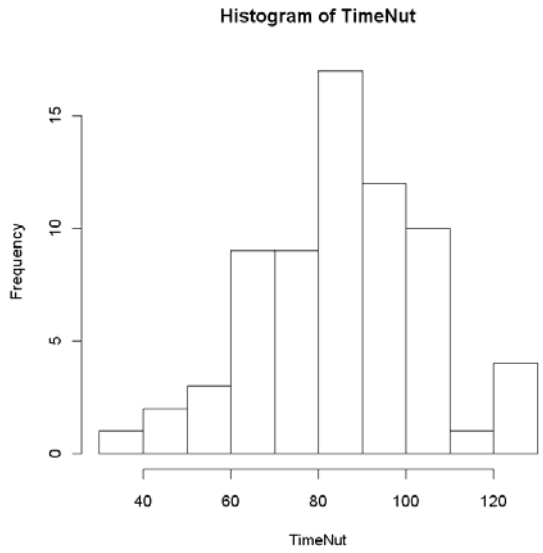
Τα δύο παραπάνω διαγράμματα προκύπτουν και με την εκτέλεση των ακόλουθων εντολών

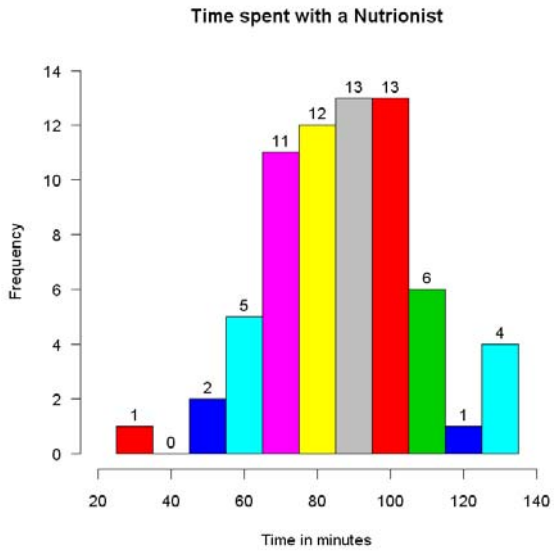
```
> pie(table(Smoke))
> dotchart(table(Smoke))
```

Η μεταβλητή TimeNut του αρχείου Infants.txt δηλώνει το χρόνο (σε λεπτά) που έχει περάσει κάθε μητέρα με διατροφολόγο και είναι ποσοτική μεταβλητή (περιέχει δεδομένα μετρήσεων (measure-

ment data)). Για την κατασκευή ενός ιστογράμματος συχνοτήτων χρησιμοποιείται η συνάρτηση `hist`. Για παράδειγμα

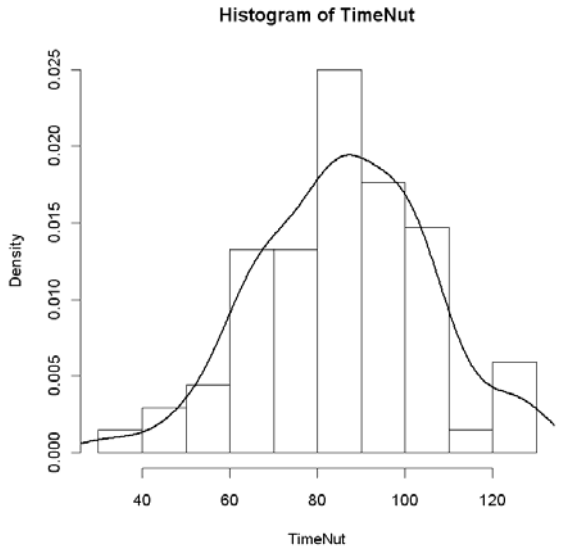
```
> class(TimeNut)
[1] "integer"
> mode(TimeNut)
[1] "numeric"
> hist(TimeNut)
> min(TimeNut)
[1] 32
> max(TimeNut)
[1] 128
> class <- seq(25,135,10)
> hist(TimeNut, breaks=class, include.lowest=TRUE, right=FALSE,
+ xlab="Time in minutes", main="Time spent with a Nutrionist",
+ xlim=c(20,140), ylim=c(0,14),labels=TRUE, col=c(2:8), las=1)
```

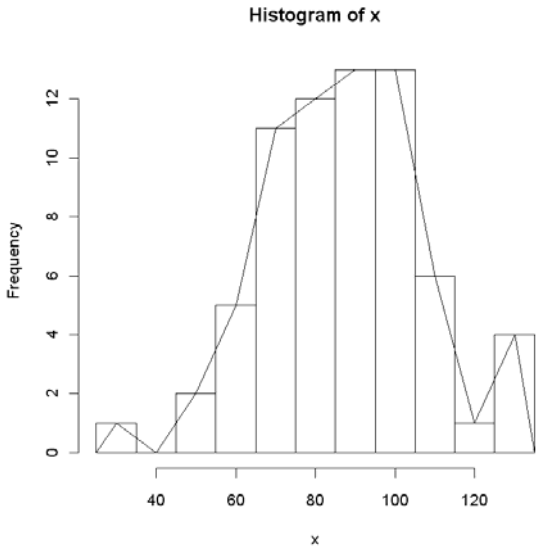




Προσέξτε ότι στο δεύτερο ιστόγραμμα κάθε κλάση περιέχει τις παρατηρήσεις που είναι ίσες με το αριστερό όριο της κλάσης και δεν περιέχει αυτές που είναι ίσες με το ανώτερο όριο (`include.lowest=TRUE, right=FALSE`). Στο ιστόγραμμα συχνοτήτων έχουμε τη δυνατότητα να προσαρμόσουμε μια καμπύλη που αποτελεί εκτίμηση της συνάρτησης πυκνότητας της μεταβλητής TimeNut μέσω της συνάρτησης `density`. Επίσης μπορούμε να δώσουμε γραφική παράσταση και του πολυγώνου συχνοτήτων χρησιμοποιώντας τη συνάρτηση `simple.freqpoly` του πακέτου UsingR. Για παράδειγμα

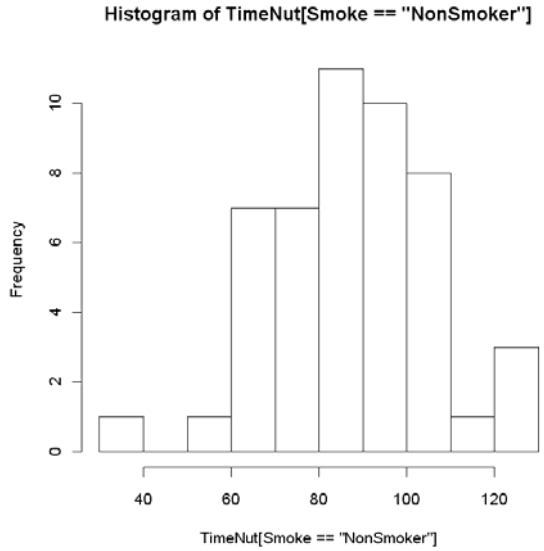
```
> hist(TimeNut,prob=TRUE)
> lines(density(TimeNut), lwd=2)
> library(UsingR)
> simple.freqpoly(TimeNut, breaks=seq(25,135,10),
+ include.lowest=TRUE, right = FALSE)
```





Η κατασκευή του ιστογράμματος της μεταβλητής TimeNut στο επίπεδο NonSmoker της μεταβλητής Smoke γίνεται ως ακολούθως

```
> hist(TimeNut[Smoke=="NonSmoker"])
```




```
  3 | 2
  3 |
  4 |
  4 | 8
  5 | 0
  5 | 7
  6 | 0034
  6 | 556779
  7 | 01122
  7 | 6779
  8 | 01223444
  8 | 5556778
  9 | 000123
  9 | 668899
 10 | 0002222
 10 | 5555
 11 | 00
 11 | 5
 12 |
 12 | 5568
```

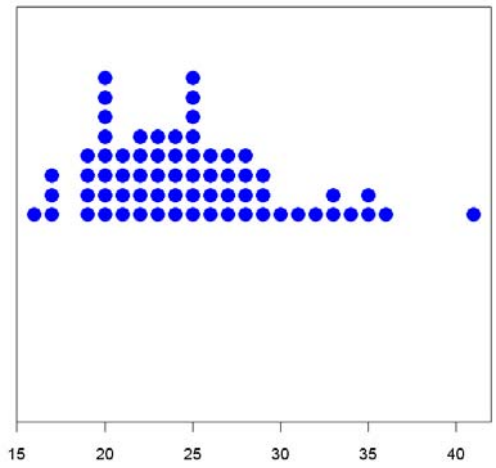
Για την κατασκευή διαγράμματος μίσχου – φύλλων (δείτε παραπάνω) χρησιμοποιείται η συνάρτηση `stem` (το όρισμα `scale` ελέγχει το μήκος του διαγράμματος). Για παράδειγμα

```
> stem(TimeNut, scale=2)
```

Ένα ιδιαίτερο γράφημα που είναι διαθέσιμο στη βασική έκδοση του R είναι το stripchart ή αλλιώς dotplot. Για τη μεταβλητή Age που δηλώνει την ηλικία της μητέρας έχουμε τα ακόλουθα γραφήματα

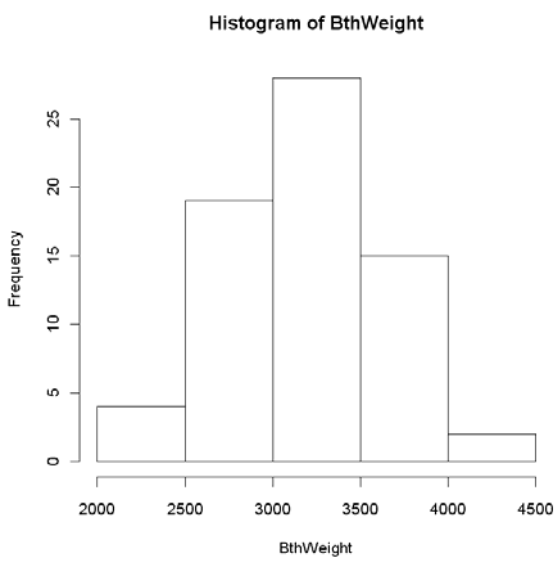
```
> stem(Age)
> stripchart(Age, method="stack",pch=16,col="blue",cex=2,offset=0.5)
```

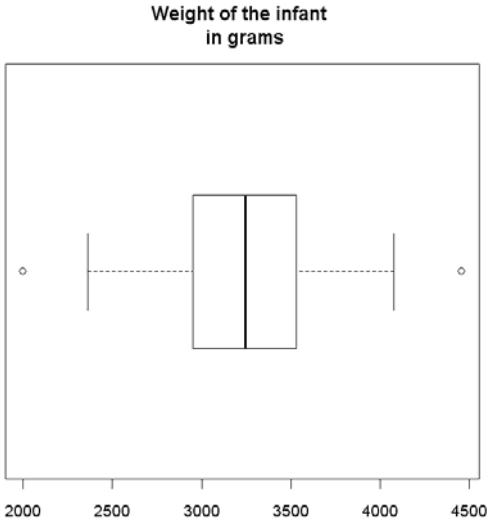
```
  1 | 67779999
  2 | 00000000111122222333344444
  2 | 55555555666677778888999
  3 | 012334
  3 | 556
  4 | 1
```



Για την κατασκευή θηκογράμματος χρησιμοποιείται η συνάρτηση `boxplot`. Για τη μεταβλητή BthWeight που δηλώνει το βάρος του βρέφους σε γραμμάρια έχουμε τα ακόλουθα γραφήματα

```
> hist(BthWeight)
> boxplot(BthWeight, main="Weight of the infant \n in grams",
+ horizontal=T)
```





Για τον υπολογισμό περιγραφικών μέτρων που αφορούν μια ποσοτική μεταβλητή μπορεί να χρησιμοποιηθεί ο Πίνακας 2.1. Για παράδειγμα

```
> mean(TimeNut)
[1] 86.17647
> median(TimeNut)
[1] 85.5
> range(TimeNut)
[1]  32 128
> diff(range(TimeNut))
[1] 96
> var(TimeNut)
[1] 374.9833
```

```
> quantile(TimeNut, 0.30, type=6)
30%
 77
> quantile(TimeNut, seq(0.1,0.9,0.1))
  10%   20%   30%   40%   50%   60%   70%   80%   90%
 63.7  69.4  77.0  82.8  85.5  90.2  98.0 102.0 106.5
> IQR(TimeNut)   #Q3-Q1
[1] 28.25
> summary(TimeNut)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   71.75   85.50   86.18  100.00  128.00
```

4.3 Ανάλυση δεδομένων: Περισσότερες μεταβλητές

4.3.1 Δύο παράγοντες

Ας θεωρήσουμε τις μεταβλητές (παράγοντες) Smoke (Levels: HeavySmoker, LightSmoker Non-Smoker) και Ethnic (Levels: Black, Hispanic, White). Για να κατασκευάσουμε τον πίνακα συνάφειας των μεταβλητών Smoke και Ethnic (two-way contingency table) εκτελούμε την εντολή `table` (απόλυτες τιμές), ή `prop.table` (σχετικές τιμές). Για παράδειγμα

```
> ct <- table(Smoke, Ethnic); ct
             Ethnic
Smoke         Black Hispanic White
  HeavySmoker     3        1     6
  LightSmoker     4        0     5
  NonSmoker      18       11    20
> prop.table(ct)                      # Εναλλακτικά  ct/sum(ct)
             Ethnic
Smoke               Black   Hispanic      White
  HeavySmoker 0.04411765 0.01470588 0.08823529
  LightSmoker 0.05882353 0.00000000 0.07352941
  NonSmoker   0.26470588 0.16176471 0.29411765
> class(ct)
[1] "table"
```

Προκειμένου να δημιουργήσουμε έναν πίνακα 2×2 με τις ίδιες πληροφορίες εκτελούμε τις εντολές

```
> x <- matrix(c(18,11,20,4,0,5,3,1,6), nrow=3, byrow=TRUE)
> dimnames(x) <-
list(c("NonSmoker","LightSmoker","HeavySmoker"),c("Black","Hispanic",
+ "White"))
> x
            Black Hispanic White
NonSmoker      18       11    20
LightSmoker     4        0     5
HeavySmoker     3        1     6
> ## Εναλλακτικός τρόπος εισαγωγής ονομάτων
> ## rownames(x)=c("NonSmoker","LightSmoker","HeavySmoker")
> ## colnames(x)=c("Black","Hispanic","White")
> class(x)
[1] "matrix"
```

Περιθώρια αθροίσματα προκύπτουν με τις συναρτήσεις `margin.table` και `addmargins`. Για παράδειγμα

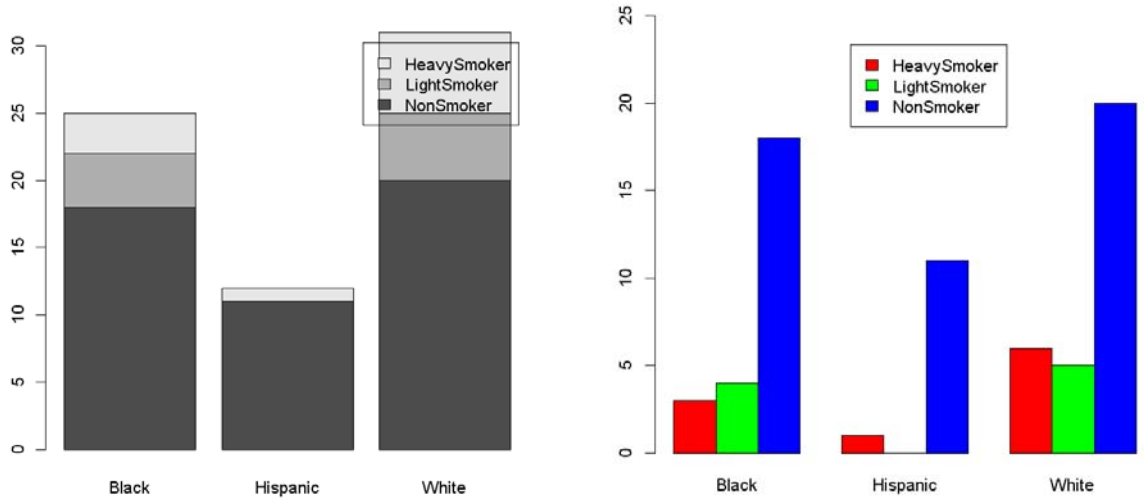
```
> margin.table(x,1)                    # Εναλλακτικά  margin.table(ct,1)
  NonSmoker LightSmoker HeavySmoker
         49           9          10
> # 1 για γραμμές, 2 για στήλες
> margin.table(x,2)                    # Εναλλακτικά  margin.table(ct,2)
   Black Hispanic    White
      25       12       31
> addmargins(x)                        # Εναλλακτικά  addmargins(ct)
            Black Hispanic White Sum
NonSmoker      18       11    20  49
LightSmoker     4        0     5   9
HeavySmoker     3        1     6  10
Sum            25       12    31  68
> prop.table(x,1)                       # Εναλλακτικά  prop.table(ct,1)
                Black  Hispanic     White
NonSmoker   0.3673469 0.2244898 0.4081633
LightSmoker 0.4444444 0.0000000 0.5555556
HeavySmoker 0.3000000 0.1000000 0.6000000
```

Για να κατασκευάσουμε πίνακες συνάφειας δύο μεταβλητών ως προς τις διαφορετικές τιμές μιας τρίτης μεταβλητής (tree-way contingency tables) εκτελούμε την εντολή `table` ή `ftable`. Για παράδειγμα

```
> table(Smoke, Ethnic, BreastFed)
, , BreastFed = No

             Ethnic
Smoke         Black Hispanic White
  HeavySmoker     3        0     2
  LightSmoker     4        0     2
  NonSmoker      10        5     6

, , BreastFed = Yes

             Ethnic
Smoke         Black Hispanic White
  HeavySmoker     0        1     4
  LightSmoker     0        0     3
  NonSmoker       8        6    14
> ftable(table(Smoke, Ethnic, BreastFed))
                    BreastFed No Yes
Smoke       Ethnic
HeavySmoker Black                3   0
            Hispanic             0   1
            White                2   4
LightSmoker Black                4   0
            Hispanic             0   0
            White                2   3
NonSmoker   Black               10   8
            Hispanic             5   6
            White                6  14
```

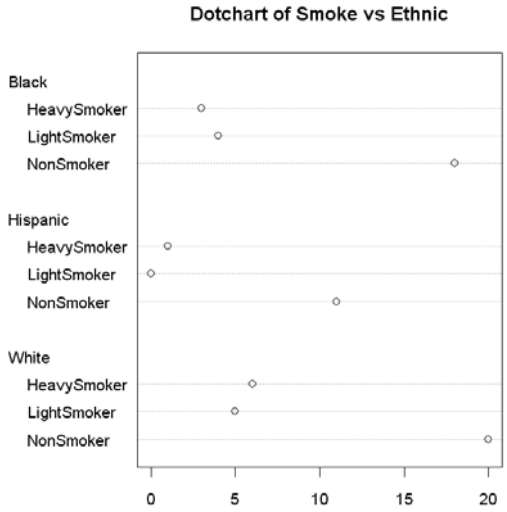
Για να κατασκευαστεί ραβδόγραμμα για κάθε στήλη ενός πίνακα χρησιμοποιούμε τη συνάρτηση `barplot()`. Για παράδειγμα

```
> barplot(x, legend.text=TRUE)
> # Εναλακτικά barplot(ct, legend.text=TRUE)
> barplot(x, beside=TRUE, col=rainbow(3), ylim=c(0,25))
> labs <- c("HeavySmoker","LightSmoker","NonSmoker")
> legend(locator(1), labs, fill=rainbow(3))
```



Με τη συνάρτηση `dotchart()` κατασκευάζεται διάγραμμα κουκκίδων για κάθε στήλη του πίνακα. Για παράδειγμα

```
> dotchart(x, main="Dotchart of Smoke vs Ethnic")
```

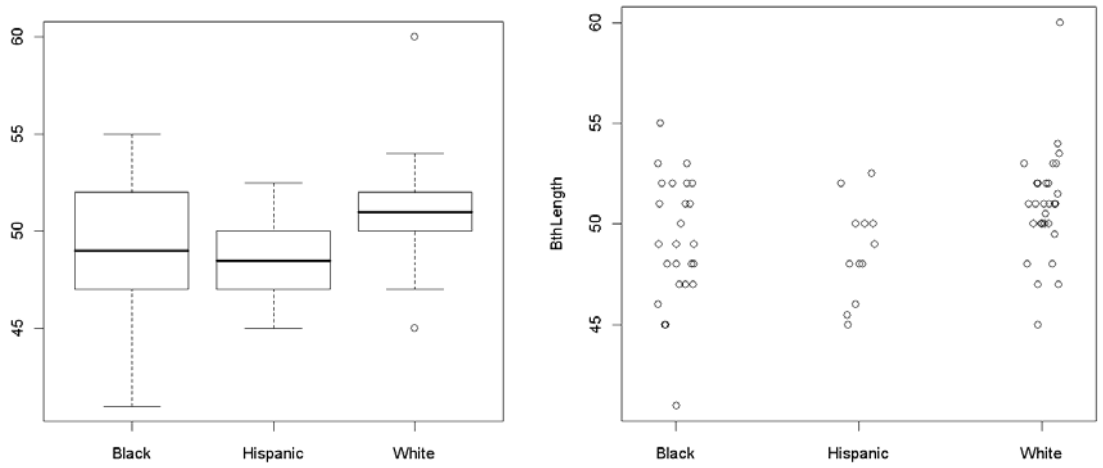


4.3.2 Ένας παράγοντας και μια μεταβλητή μετρήσεων

Ας θεωρήσουμε τις μεταβλητές Ethnic και BthLength (μήκος του βρέφους σε εκατοστά) που είναι, αντίστοιχα, κατηγορική μεταβλητή και μεταβλητή μετρήσεων, αντίστοιχα. Για τη μελέτη αυτού του

τύπου συνδυασμού μεταβλητών δίνουμε το κάτωθι παράδειγμα

```
> plot(Ethnic, BthLength)             # Θηκογράμματα ανά Ethnic
> stripchart(BthLength~Ethnic, pch=1, method="jitter", vertical=TRUE)
> tapply(BthLength,Ethnic,summary)

$Black
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.00   47.00   49.00   49.16   52.00   55.00
$Hispanic
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  45.00   47.50   48.50   48.67   50.00   52.50
$White
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   45.0    50.0    51.0    50.9    52.0    60.0
```

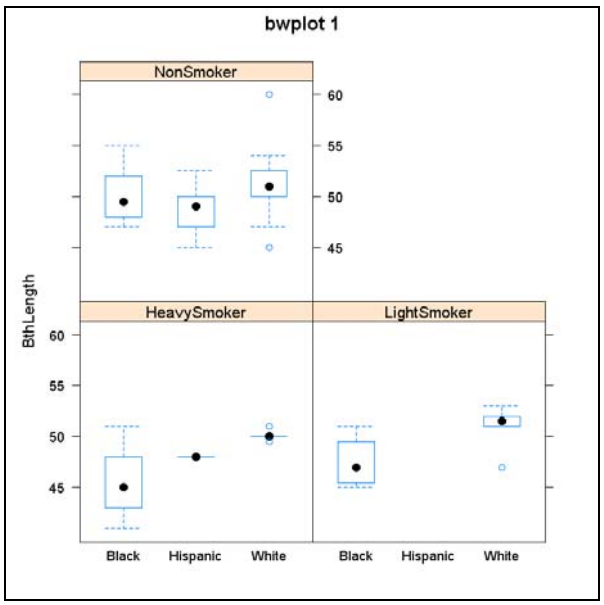


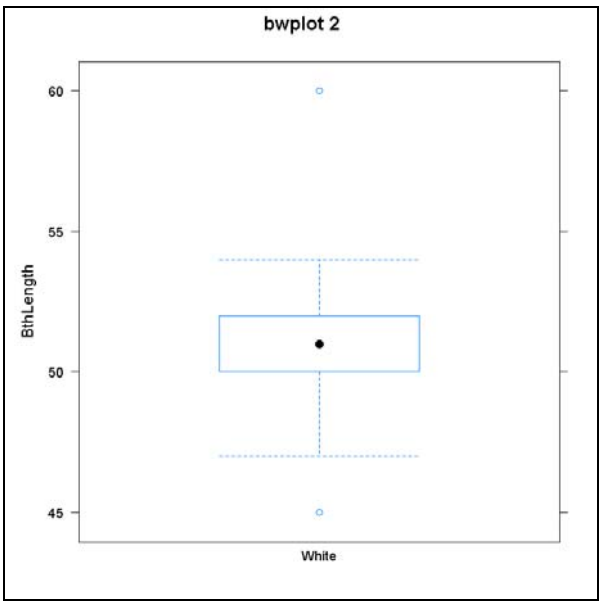
Ενδιαφέρουσες εναλλακτικές γραφικές παραστάσεις σε πάνελ αλλά και ευκολία στο χειρισμό των μεταβλητών προσφέρει το πακέτο lattice (Cleveland's Trellis graphics concepts). Οι συναρτήσεις του πακέτου χρησιμοποιούν εξαρτήσεις μεταξύ των μεταβλητών της μορφής

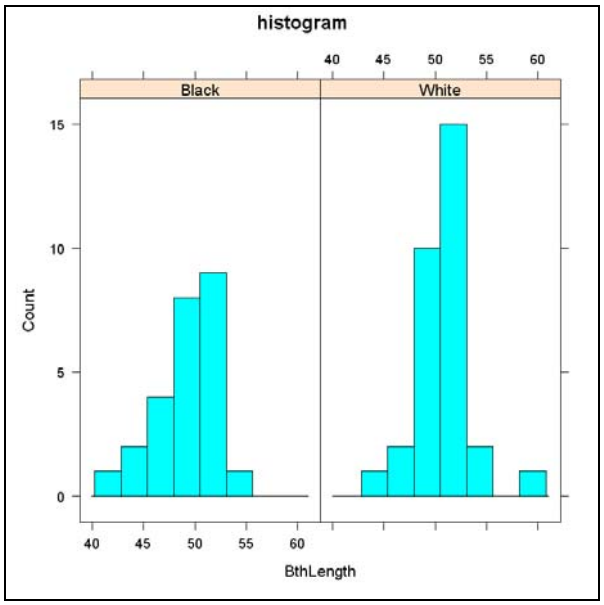
$$y \sim x \mid z \quad (\text{response} \sim \text{predictor} \mid \text{condition}) \quad \text{ή} \quad \sim x \mid z$$

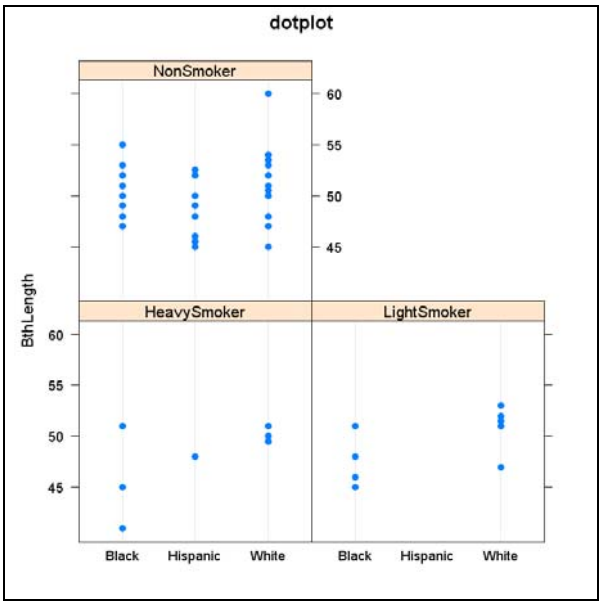
Για παράδειγμα δίνεται το ακόλουθο σετ τεσσάρων γραφημάτων (γίνεται χρήση της συνάρτησης-ορίσματος `subset`).

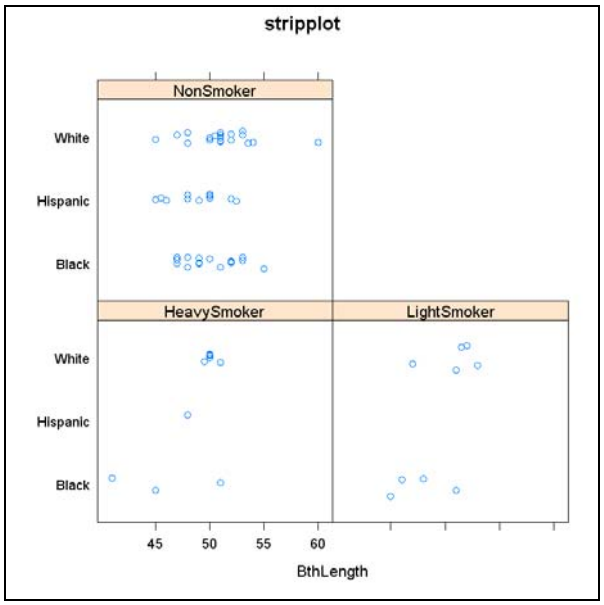
```
> library(lattice)
> bwplot(BthLength~Ethnic|Smoke, main="bwplot 1")
> bwplot(BthLength~Ethnic,subset=(Ethnic=="White"), main="bwplot 2")
> histogram( ~BthLength | Ethnic,
+ subset=(Ethnic=="Black")|(Ethnic=="White"), type="count",
+ nint=8, main="histogram")
> dotplot(BthLength~Ethnic|Smoke, main="dotplot")
> stripplot(Ethnic~BthLength|Smoke, jitter=TRUE, main="stripplot")
> xyplot(BthLength~Age|Smoke, main="xyplot")
```

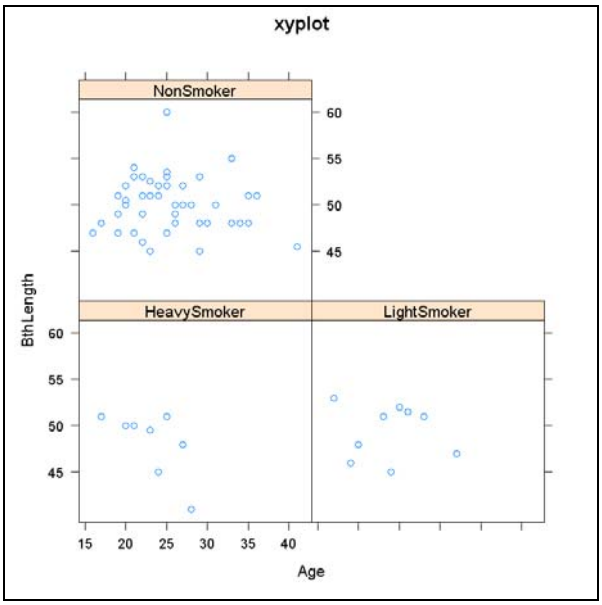

bwplot 1
NonSmoker
HeavySmoker
LightSmoker
BthLength
Black
Hispanic
White


bwplot 2
BthLength
White


histogram
Black
White
Count
BthLength


dotplot
NonSmoker
HeavySmoker
LightSmoker
BthLength
Black
Hispanic
White


stripplot
NonSmoker
HeavySmoker
LightSmoker
White
Hispanic
Black
BthLength


xyplot
NonSmoker
HeavySmoker
LightSmoker
BthLength
Age

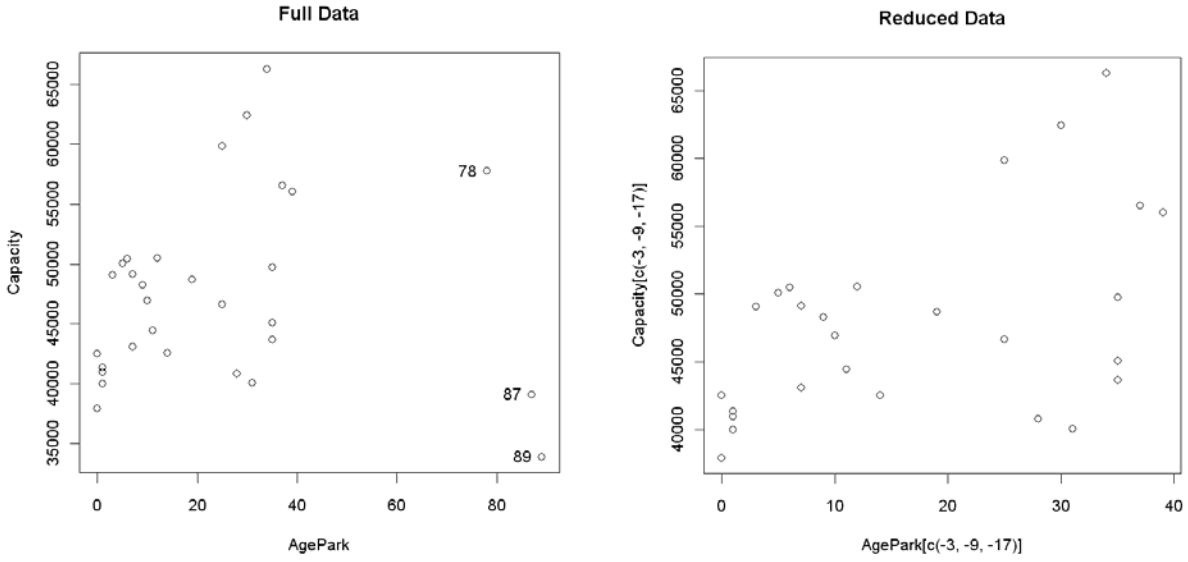
4.3.3 Δύο ή περισσότερες μεταβλητές μετρήσεων

Το αρχείο BallParkData.txt που βρίσκεται στον κατάλογο εργασίας του R περιέχει δεδομένα που αφορούν 30 ομάδες μπέιζμπολ που αγωνίστηκαν την περίοδο 2001. Η μεταβλητή League δηλώνει την κατηγορία που αγωνίζεται η ομάδα (American, National), η μεταβλητή ParkBlt δηλώνει τη χρονολογία που χτίστηκε το στάδιο κάθε ομάδας, η μεταβλητή Capacity δηλώνει τη χωρητικότητα του σταδίου, η μεταβλητή Attend δηλώνει το μέσο όρο των φιλάθλων που παρακολούθησαν τα παιχνίδια της ομάδας το έτος 2001, και η μεταβλητή WinPct το ποσοστό των νικών κάθε ομάδας. Για την εισαγωγή των δεδομένων στο R εκτελούμε τις εντολές

```
> b <-read.table("BallParkData.txt", header=TRUE)
> b
                    Team   League ParkBlt Capacity Attend WinPct
1          Anaheim-Angels American    1966    45050  24708  0.463
2       Baltimore-Orioles American    1992    48262  38686  0.391
.....................................................................
30  San-Francisco-Giants National    2000    41341  40877  0.556
> attach(b)
> class(b)
[1] "data.frame"
> names(b)
[1] "Team"     "League"   "ParkBlt"  "Capacity" "Attend"   "WinPct"
```

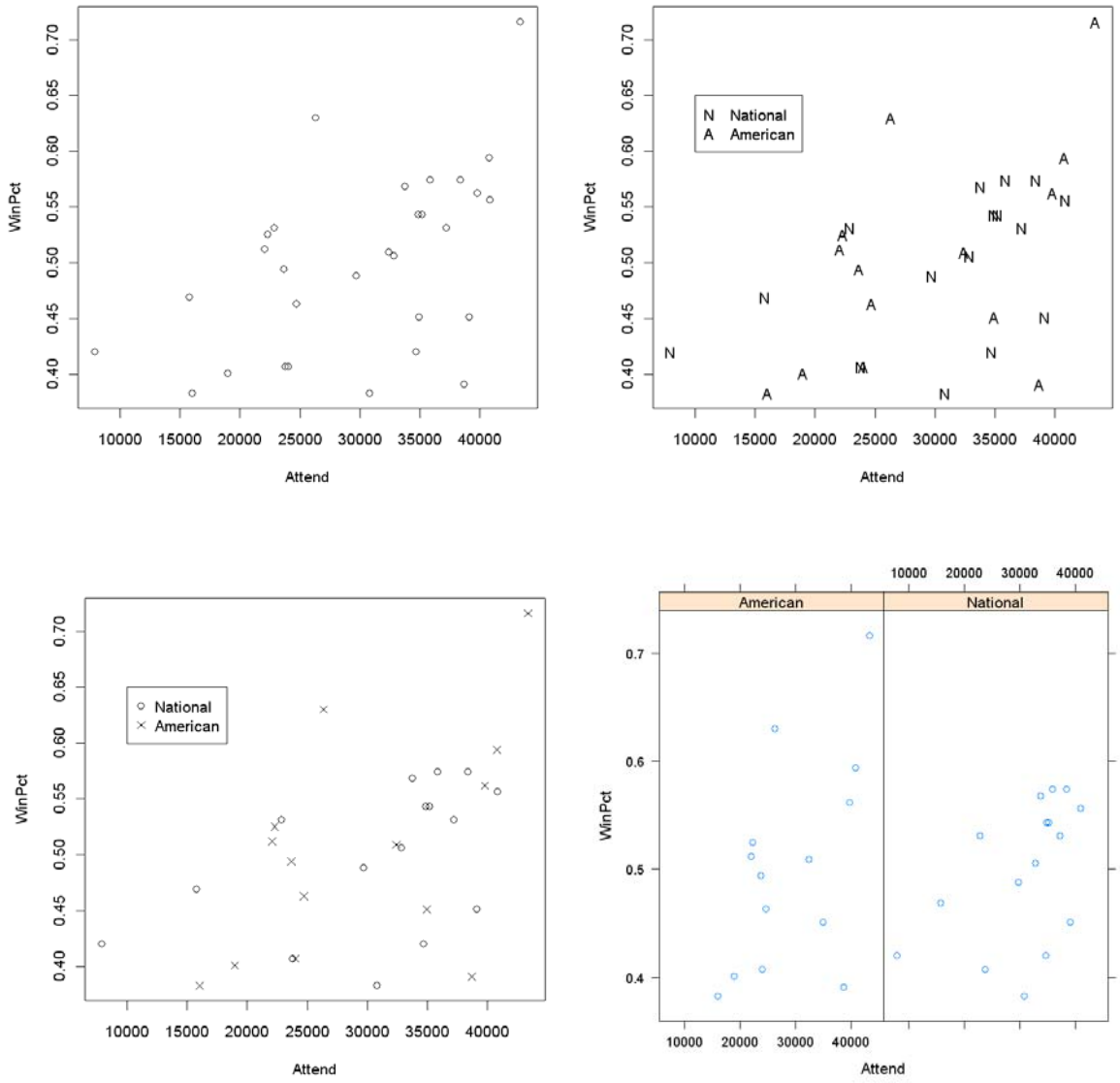
Για να εξετάσουμε αν υπάρχει κάποιου είδους συσχέτιση μεταξύ της ηλικίας και της χωρητικότητας των σταδίων κατασκευάζεται ένα διάγραμμα διασποράς απ'όπου αφαιρούνται στη συνέχεια οι ακραίες παρατηρήσεις.

```
> AgePark <- 2001-ParkBlt
> plot(AgePark, Capacity, main="Full Data")
> identify(AgePark, Capacity, labels=AgePark)
[1]  3  9 17
> plot(AgePark[c(-3,-9,-17)],Capacity[c(-3,-9,-17)], main="Reduced Data")
```



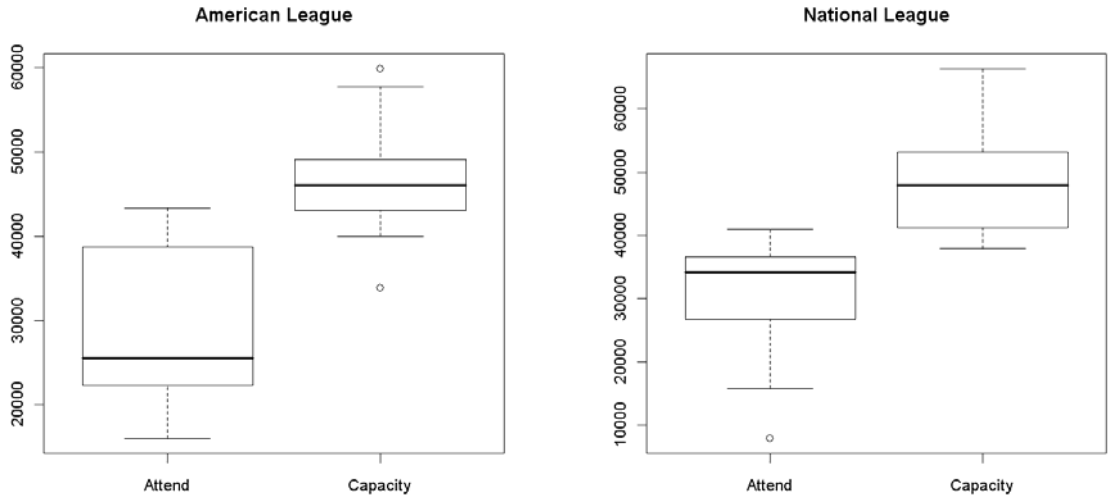
Για να σχεδιάσουμε ένα διάγραμμα διασποράς των μεταβλητών Attend και WinPct με διάκριση των σημείων του γραφήματος ως προς μια τρίτη μεταβλητή, την League, δίνουμε το ακόλουθο παράδειγμα (γίνεται χρήση της συνάρτησης `ifelse`).

```
> plot(Attend,WinPct)
> plot(Attend,WinPct,pch=as.character(League))
> legend(10000,0.65,legend=c("National","American"), pch=c("N", "A"))
> plot(Attend,WinPct,pch=ifelse(League=="National",1,4))
> legend(10000,0.65,legend=c("National","American"), pch=c(1,4))
> library(lattice)
> xyplot(WinPct~Attend|League)
```



Για να κατασκευάσουμε θηκογράμματα των μεταβλητώνAttend (5η στήλη του data.frame b) και Capacity (4η στήλη του data.frame b) για κάθε μια από τις δύο βασικές κατηγορίες (League) εργαζόμαστε ως ακολούθως.

```
> LeagueA <- subset(b,subset=League=="American", select=c(5,4))
> boxplot(LeagueA,main="American League")
> LeagueN <- subset(b,subset=League=="National", select=c(5,4))
> boxplot(LeagueN,main="National League")
```



Για να δώσουμε τον πίνακα συσχέτισης των μεταβλητών Attend, Capacity και WinPct εκτελούμε τη συνάρτηση `cor`. Για τα βασικά περιγραφικά κάθε μεταβλητής εκτελούμε τη συνάρτηση `summary`.

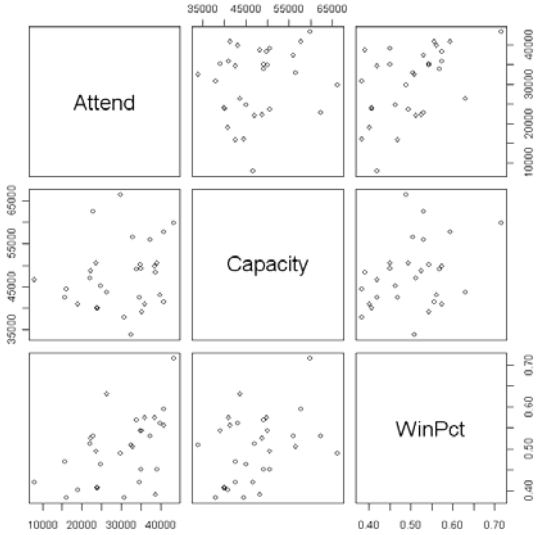
```
> y <- cbind(Attend,Capacity,WinPct)
> is.matrix(y)
[1] TRUE
> cor(y)
            Attend  Capacity    WinPct
Attend   1.0000000 0.1891696 0.4985423
Capacity 0.1891696 1.0000000 0.3586474
WinPct   0.4985423 0.3586474 1.0000000
> summary(y)
     Attend         Capacity         WinPct
 Min.   : 7935   Min.   :33871   Min.   :0.3830
 1st Qu.:23716   1st Qu.:41631   1st Qu.:0.4278
 Median :32616   Median :46782   Median :0.5075
 Mean   :30062   Mean   :47451   Mean   :0.5001
 3rd Qu.:36907   3rd Qu.:50352   3rd Qu.:0.5527
 Max.   :43362   Max.   :66307   Max.   :0.7160
```

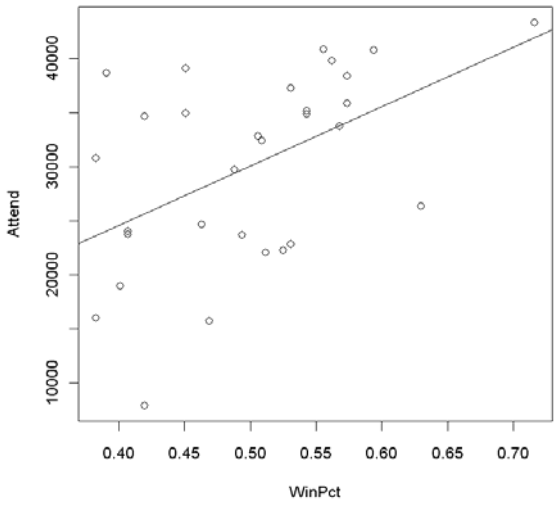
Για να δώσουμε διαγράμματα διασποράς των μεταβλητών Attend, Capacity και WinPct ανά δύο χρησιμοποιούμε κατάλληλα την εντολή `plot` (ή την εντολή `pairs`). Για να σχεδιάσουμε την ευθεία ελάχιστων τετραγώνων

$$\text{Attend} = b_0 + b_1 \text{WinPct}$$

χρησιμοποιούμε τη συνάρτηση `lm`.

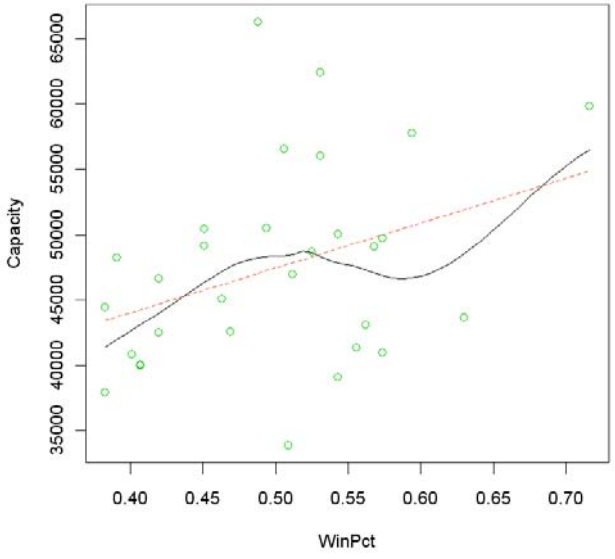
```
> plot(b[c(5,4,6)])    # εναλλακτικά > pairs(b[c(5,4,6)])
> lm(Attend~WinPct)

Call:
lm(formula = Attend ~ WinPct)

Coefficients:
(Intercept)       WinPct
       2560        54997
> lr <- lm(Attend~WinPct)
> plot(WinPct,Attend)
> abline(lr)
```





Κλείνοντας σημειώνουμε ότι υπάρχουν αρκετές μέθοδοι σχεδίασης καμπυλών που μπορούν να προσαρμοστούν στα δεδομένα μας. Για παράδειγμα

```
> scatter.smooth(WinPct,Capacity,col="green3")
> lines(smooth.spline(WinPct,Capacity),lty=2, col="tomato2")
```



4.4 Σύνοψη εντολών Κεφαλαίου 4

```
addmargins, as.character, attach
barplot, boxplot, bwplot
data, density, diff, dotchart, dotplot, dump
ftable
hist
ifelse, identify
legend, lm, locator
margin.table
names
pie, plot, prop.table
read.table, read.spss
scatter.smooth, smooth.spline, simple.freqpoly, source, stem, stripchart,
simple.freqpoly, stripplot, subset
table, tapply
write.table
xyplot
```

ΚΕΦΑΛΑΙΟ 5

Στοιχεία Πιθανοτήτων

5.1 Παραγωγή τυχαίων αριθμών

Η παραγωγή τυχαίων αριθμών στο R γίνεται με τη συνάρτηση `sample`. Για παράδειγμα

```
> k1 <- 1:20
> sample(k1,size=10,replace=TRUE)       # Με επανατοποθέτηση
 [1]  8 13  3  1 10 11  9  1  6 18
> sample(k1,size=5,replace=FALSE)       # Χωρίς επανατοποθέτηση
[1] 15  9 11  5  7
```

Στο παραπάνω παράδειγμα κάθε αριθμός που παράγεται έχει την ίδια πιθανότητα να επιλεγεί από τους αριθμούς που αποτελούν τον πληθυσμό. Αν θέλουμε να υπάρχει διαφορετική πιθανότητα χρησιμοποιούμε το όρισμα `prob`. Για παράδειγμα

```
> k <- 0:4
> p <- c(1,1,2,3,4)/11
> sample(k,size=15,prob=p,replace=TRUE)
 [1] 3 3 1 0 4 4 4 2 1 3 1 0 0 4 3
> sample(0:1,size=10,replace=TRUE,prob=c(0.32,1-0.32))  # Τυχαίο δείγμα
 [1] 0 1 1 1 0 1 1 0 0 0                                # από Bernoulli
```

Για να παραχθεί το ίδιο σύνολο τυχαίων αριθμών σε διαδοχικές εκτελέσεις της συνάρτησης `sample` χρησιμοποιείται η συνάρτηση `set.seed`.

```
> r <- 1:20
> set.seed(136); r1 <- sample(r,size=8,replace=TRUE);r1
[1]  7  8  9 20  9 11 20 12
> set.seed(136); r2 <- sample(r,size=8,replace=TRUE);r2
[1]  7  8  9 20  9 11 20 12
```

5.2 Κατανομές

Με το R μπορούμε να υπολογίζουμε πιθανότητες, αθροιστικές πιθανότητες, ποσοστιαία σημεία, κτλ., για διάφορες κατανομές. Εκτελώντας την εντολή

```
> help.search("distribution")
```

εμφανίζονται σε ένα παράθυρο οι κατανομές που είναι διαθέσιμες (οι περισσότερες φορτώνονται από το πακέτο stats). Πληροφορίες για τις κατανομές, όπως η ακριβή μορφή της συνάρτησης πυκνότητας ή πιθανότητας, μπορούν να βρεθούν εκτελώντας την εντολή

`?Όνομα_κατανομής(όνομα_πακέτου).`

Για παράδειγμα

```
> ?Cauchy(stats)
```

Οι βασικότερες κατανομές και οι παράμετροί τους δίνονται στον ακόλουθο πίνακα.

Πίνακας 5.1: Κατανομές

Κατανομή	**R Όνομα (Rname)**	**Παράμετροι**
Beta	`beta`	shape1, shape2, ncp
Binomial	`binom`	size, prob
Cauchy	`cauchy`	location, scale
Chisquare	`chisq`	df, ncp
Exponential	`exp`	rate
FDist	`f`	df1, df2, ncp
GammaDist	`gamma`	shape, scale
Geometric	`geom`	prob
Hypergeometric	`hyper`	m, n, k
Lognormal	`lnorm`	meanlog, sdlog
Logistic	`logis`	location, scale
NegBinomial	`nbinom`	size, prob
Normal	`norm`	mean, sd
Poisson	`pois`	lambda
TDist	`t`	df, ncp
Uniform	`unif`	min, max
Weibull	`weibull`	shape, scale
Wilcoxon	`wilcox`	m, n
Multinomial	`multinom`	size, prob

Βάζοντας τα προθέματα `d`, `p`, `q` και `r` πριν από το R όνομα (Rname) της κατανομής προκύπτει, αντίστοιχα, η συνάρτηση πυκνότητας ή πιθανότητας (σ.π.), η συνάρτηση κατανομής (σ.κ.), ποσοστιαία σημεία και τυχαίοι αριθμοί της κατανομής. Πιο συγκεκριμένα

- `dRname(x, ...)` `# Υπολογισμός της σ.π. στο x`
- `pRname(q, ...)` `# Υπολογισμός της σ.κ. στο q`
- `qRname(p, ...)` `# Υπολογισμός του p-ποσοστιαίου σημείου`
- `rRname(n, ...)` `# Παραγωγή n τυχαίων αριθμών`

5.2.1 Διωνυμική κατανομή

Η συνάρτηση πιθανότητας της διωνυμικής κατανομής με παραμέτρους n και p (συμβ. $B(n, p)$) δίνεται από τον τύπο

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x},\quad x=0,1,2,...,n\,.$$

Πληροφορίες για τη διωνυμική κατανομή προκύπτουν με την εκτέλεση της εντολής "`> ?Binomial(stats)`". Για τη διωνυμική κατανομή $B(10,0.5)$ δίνουμε το ακόλουθο παράδειγμα

```
> n <- 10; p <- 1/2; x <- 5
> choose(n,x)*p^x*(1-p)^(n-x)                # P(X=5)
[1] 0.2460938
> dbinom(x,size=n,prob=p)                    # P(X=5)
[1] 0.2460938
> sum(dbinom(0:x,size=n,prob=p))             # F(5)
[1] 0.6230469
> pbinom(x,size=n,prob=p)                    # F(5)
[1] 0.6230469
> 1-pbinom(x,size=n,prob=p)                  # 1-F(5)=P(X>5)
[1] 0.3769531
> pbinom(x,size=n,prob=p, lower.tail=FALSE)  # P(X>5)
[1] 0.3769531
> sum(dbinom((x+1):10,size=n,prob=p))        # P(X>5))
[1] 0.3769531
```

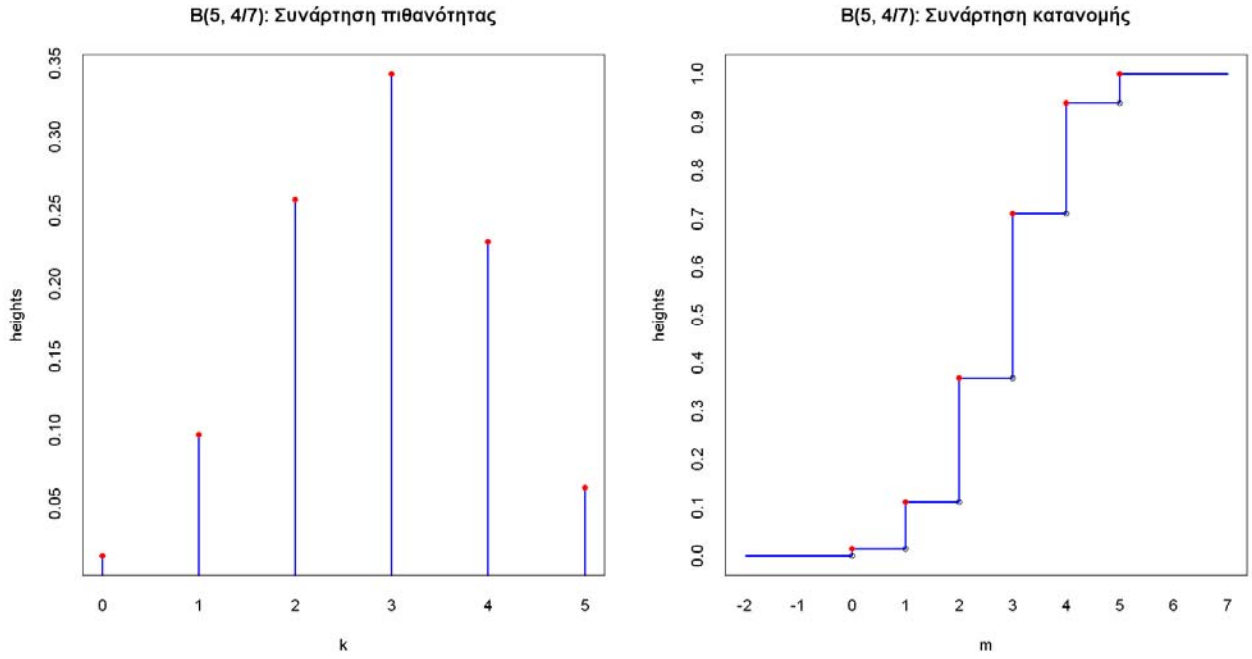
Η συνάρτηση πιθανότητας της διωνυμικής κατανομής $B(21,0.5)$ προκύπτει ως ακολούθως

```
> x1 <- 0:10
> x2 <- 11:21
> p1 <- round(dbinom(0:10, 21, 0.5), digits=8)
> p2 <- round(dbinom(11:21, 21, 0.5), digits=8)
> df <- data.frame(x1, p1, x2, p2)
> colnames(df) <- c("x", "P(X=x)","x", "P(X=x)")
> df
    x     P(X=x)  x     P(X=x)
1   0 0.00000048 11 0.16818810
2   1 0.00001001 12 0.14015675
3   2 0.00010014 13 0.09703159
4   3 0.00063419 14 0.05544662
5   4 0.00285387 15 0.02587509
6   5 0.00970316 16 0.00970316
7   6 0.02587509 17 0.00285387
8   7 0.05544662 18 0.00063419
9   8 0.09703159 19 0.00010014
10  9 0.14015675 20 0.00001001
11 10 0.16818810 21 0.00000048
```

Η γραφική παράσταση της συνάρτησης πιθανότητας και της συνάρτησης κατανομής της διωνυμικής κατανομής $B(5,4/7)$ προκύπτει ως ακολούθως

```
> par(mfcol=c(1,2))
> n <- 5; p <- 4/7; k <-0:n
> # Συνάρτηση πιθανότητας
> heights <- dbinom(0:n,size=n,prob=p)
> plot(k,heights,type="h",tck=0, lwd=2, col="blue", main="B(5, 4/7):
+ Συνάρτηση πιθανότητας")
```

```
> points(k,heights,pch=16,cex=0.8, col="red")
> # Συνάρτηση κατανομής
> m <- (-2):(n+2)
> heights <- c(0,0,pbinom(0:n,size=n,prob=p),1,1)
> plot(m,heights,type="s",lab=c(10,10,7),tck=0, lwd=2, col="blue",
+ main="B(5, 4/7): Συνάρτηση κατανομής")
> # Εισαγωγή των σημείων ο
> m1 <- 0:n
> heights1<- c(0,pbinom(0:(n-1),size=n,prob=p))
> points(m1,heights1,pch=1, cex=0.8)
> # Εισαγωγή των σημείων •
> m2 <- 0:n
> heights2<- c(pbinom(0:n,size=n,prob=p))
> points(m2,heights2, pch=16, cex=0.8, col="red")
```



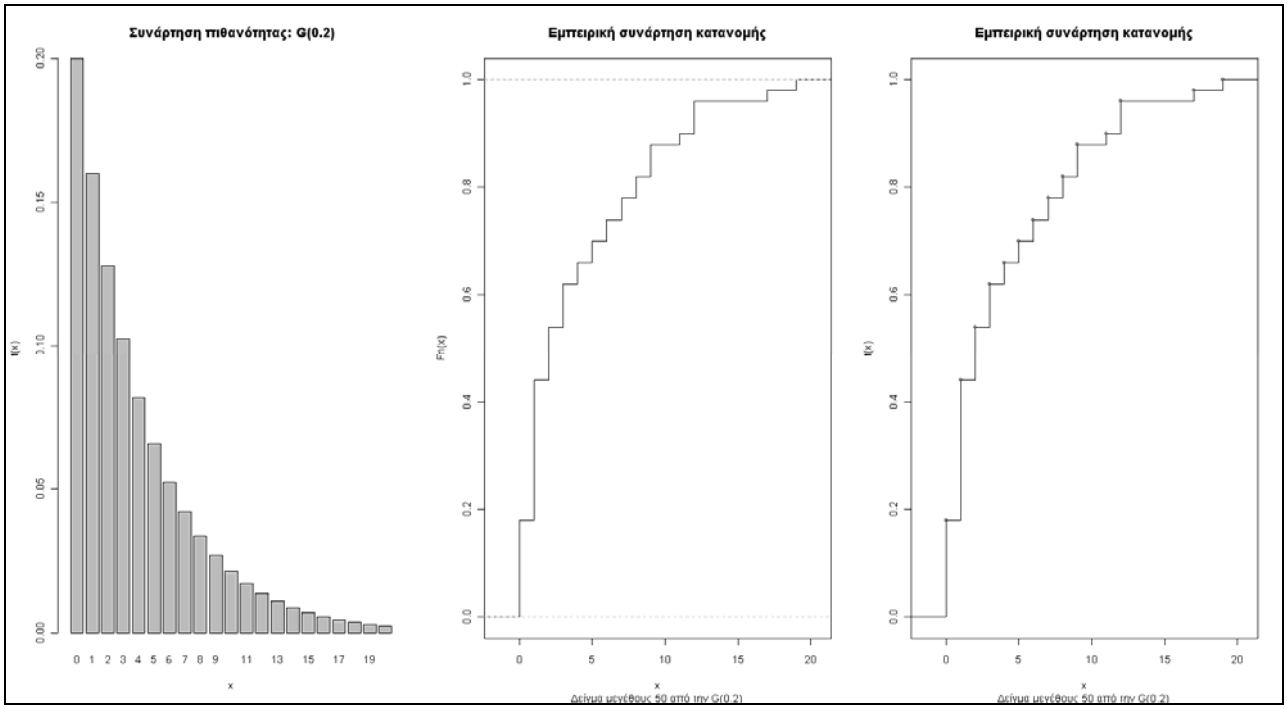
5.2.2 Γεωμετρική κατανομή

Η συνάρτηση πιθανότητας της γεωμετρικής κατανομής με παράμετρο p (συμβ. $G(p)$) δίνεται από τον τύπο

$$f(x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots .$$

Πληροφορίες για τη γεωμετρική κατανομή προκύπτουν με την εκτέλεση της εντολής "`> ?Geometric(stats)`". Για τη γεωμετρική κατανομή $G(0.2)$ δίνουμε ακολούθως τη γραφική παράσταση της συνάρτησης πιθανότητάς της $f(x)$. Επίσης δίνουμε ένα πίνακα συχνοτήτων 50 τυχαίων αριθμών από την $G(0.2)$ και την αντίστοιχη εμπειρική συνάρτηση κατανομής με χρήση των συναρτήσεων `ecdf` και της `plot.stepfun`.

```
> par(mfrow=c(1,3))
> f <- dgeom(0:20, 0.2)
> barplot(f, names=as.character(0:20), xlab="x", ylab="f(x)", main="
+ Συνάρτηση πιθανότητας: G(0.2)")
> rg1 <- rgeom(50, 0.2);rg1
> rg2 <- table(rg1);rg2
rg1
 0  1  2  3  4  5  6  7  8  9 10
11  9  4  4  6  6  2  2  2  3  1
> plot(ecdf(rg1), verticals=TRUE, do.points=FALSE, main="Εμπειρική
+ συνάρτηση κατανομής", sub="Δείγμα μεγέθους 50 από την G(0.2)")
> plot.stepfun(rg1, main="Εμπειρική συνάρτηση κατανομής", sub="Δείγμα
+ μεγέθους 50 από την G(0.2)")
```



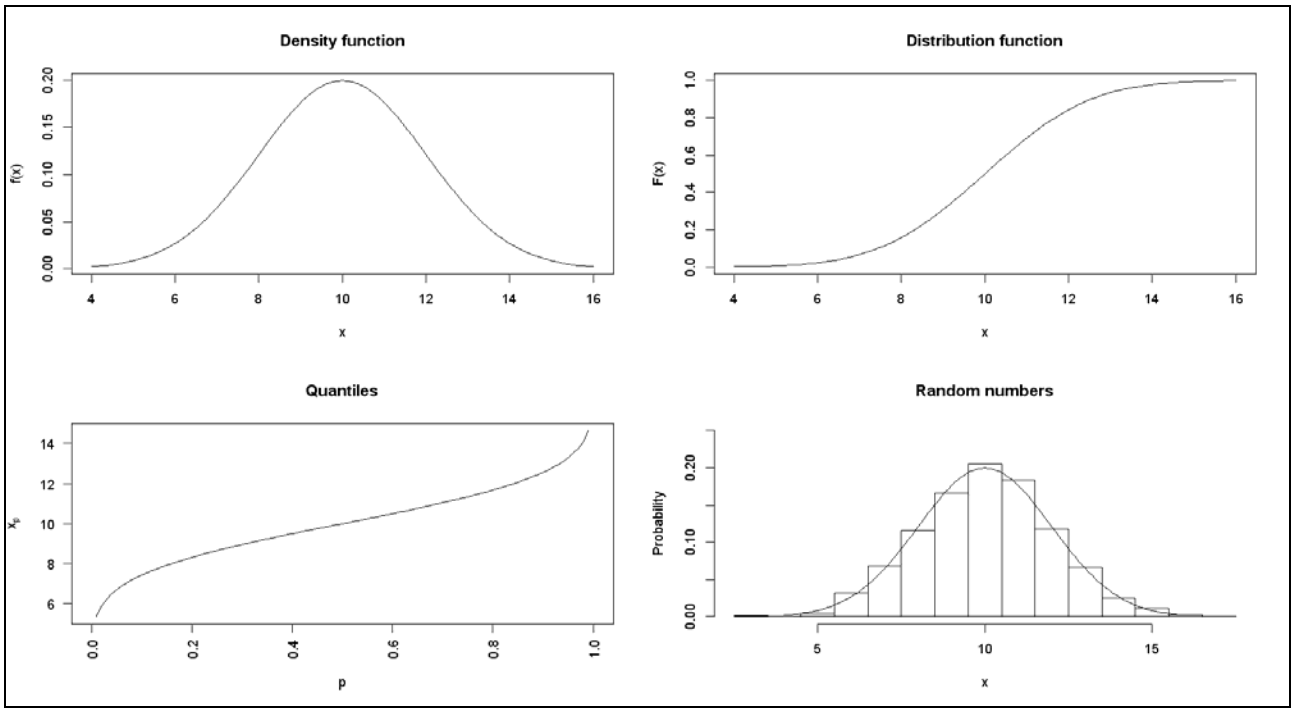
5.2.3 Κανονική κατανομή

Η κανονική κατανομή με παραμέτρους μ και σ^2 (συμβ. $N(\mu,\sigma^2)$) δίνεται από τον τύπο

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),\quad -\infty<x<\infty.$$

Πληροφορίες για την κανονική κατανομή προκύπτουν με την εκτέλεση της εντολής "`> ?Normal(stats)`". Για την κανονική κατανομή $N(10,4)$ δίνουμε γραφική παράσταση της συνάρτησης πυκνότητάς της $f(x)$, της συνάρτησης κατανομής της $F(x)$ και των ποσοστιαίων σημείων της x_p. Επίσης δίνουμε ένα ιστόγραμμα πυκνότητας σχετικών συχνοτήτων 1000 τυχαίων αριθμών από τη $N(10,4)$ στο οποίο έχει σχεδιαστεί η σ.π. $f(x)$ της $N(10,4)$.

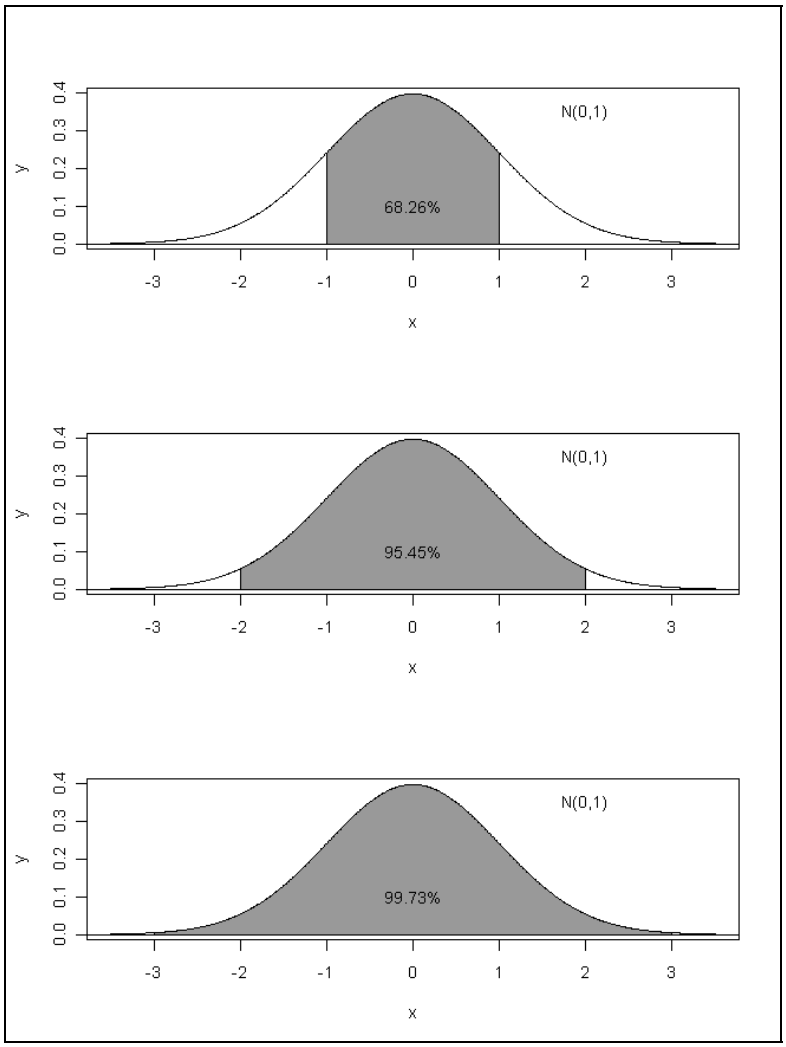
```
> par(mfrow=c(2,2))
> curve(dnorm(x, mean = 10, sd = 2),from=4,to=16, xlab="x", ylab="f(x)",
+ main="Density function")
> curve(pnorm(x, mean = 10, sd = 2),from=4,to=16, xlab="x", ylab="F(x)",
+ main="Distribution function")
> curve(qnorm(x, mean = 10, sd = 2),from=0,to=1, xlab="p",
+ ylab=expression(x[p]), las=2, main="Quantiles")
> y <- rnorm(1000, mean = 10, sd = 2)
> hist(y, breaks=2.5:17.5, prob=TRUE, ylim=c(0,0.25), xlab="x",
+ ylab="Probability", main="Random numbers")
> lines(seq(4,16,0.1),dnorm(seq(4,16,0.1), mean = 10, sd = 2))
```



Στο ακόλουθο σχήμα απεικονίζονται οι πιθανότητες $P(-k \le Z \le k)$ για $k = 1,2,3$, όπου $Z \sim N(0,1)$.

```
> par(mfrow=c(3,1))                    # P(-1 ≤ Z ≤ 1)
> x <- seq(-3.5,3.5,0.01)
> y <-dnorm(x)
> plot(x,y,type="l")
> x1 <-c(-1,c(x[x>=-1 & x<=1]),1)
> y1 <-c(0,c(y[x>=-1 & x<=1]),0)
> polygon(x1,y1,col="gray60")
> pnorm(1)-pnorm(-1)
[1] 0.6827
> options("digits"=4)
> text(0,0.1,label="68.26%")
> text(2,0.35,label="N(0,1)")
> abline(h=0)
>
> x <- seq(-3.5,3.5,0.01)              # P(-2 ≤ Z ≤ 2)
> y <-dnorm(x)
> plot(x,y,type="l")
> x1 <-c(-2,c(x[x>=-2 & x<=2]),2)
```

```
> y1 <-c(0,c(y[x>=-2 & x<=2]),0)
> polygon(x1,y1,col="gray60")
> pnorm(2)-pnorm(-2)
[1] 0.9545
> options("digits"=4)
> text(0,0.1,label="95.45%")
> text(2,0.35,label="N(0,1)")
> abline(h=0)
>
> x <- seq(-3.5,3.5,0.01)                    # P(−3 ≤ Z ≤ 3)
> y <-dnorm(x)
> plot(x,y,type="l")
> x1 <-c(-3,c(x[x>=-3 & x<=3]),3)
> y1 <-c(0,c(y[x>=-3 & x<=3]),0)
> polygon(x1,y1,col="gray60")
> pnorm(3)-pnorm(-3)
[1] 0.9973
> options("digits"=4)
> text(0,0.1,label="99.73%")
> text(2,0.35,label="N(0,1)")
> abline(h=0)
```



Για την κανονική κατανομή $N(100,100)$ οι πιθανότητες $P(\mu-k\sigma\le X\le\mu+k\sigma)=P(-k\le Z\le k)$ για $k=1,2,3$ μπορούν να προκύψουν προσεγγιστικά ως ακολούθως

```
> options("digits"=6)
> mu <- 100; sigma <- 10; size <- 100000
> res <- rnorm(size, mean=mu, sd=sigma)
> k1 <- 1; k2 <- 2; k3 <- 3
> sum(res>mu-k1*sigma & res<mu+k1*sigma)/size
[1] 0.68312
> sum(res>mu-k2*sigma & res<mu+k2*sigma)/size
[1] 0.95512
> sum(res>mu-k3*sigma & res<mu+k3*sigma)/size
[1] 0.99736
```

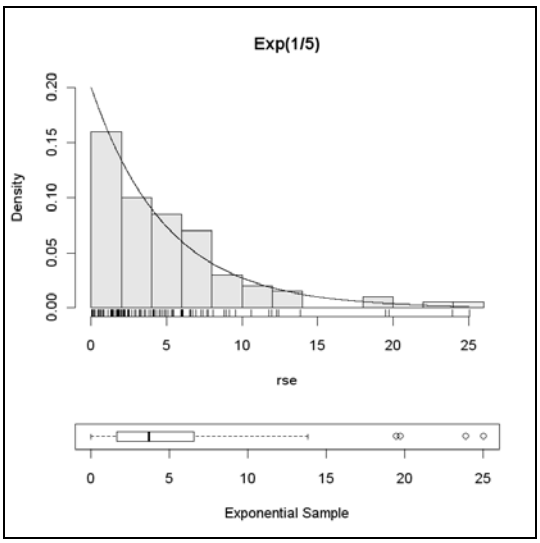
5.2.4 Εκθετική κατανομή

Για την εκθετική κατανομή με συνάρτηση πυκνότητας

$$f(x)=\lambda e^{-\lambda x},\quad x\ge 0.$$

Πληροφορίες για την εκθετική κατανομή προκύπτουν με την εκτέλεση της εντολής "`> ?Exponential(stats)`". Στο ακόλουθο παράδειγμα δίνουμε μια γραφική παράσταση που περιέχει ένα θηκόγραμμα 100 τυχαίων αριθμών από την εκθετική κατανομή με $\lambda=4$ και το αντίστοιχο ιστόγραμμα πυκνότητας σχετικής συχνότητας στο οποίο έχει επικαθίσει η συνάρτηση πυκνότητας πιθανότητας της κατανομής.

```
> rse <- rexp(100, rate=1/5)      #  λ=1/5
> par(fig=c(0,1,0,.35))
> boxplot(rse, horizontal=TRUE, xlab="Exponential Sample")
> par(fig=c(0,1,0.25,1), new=TRUE)
> # ΑΡΧΗ εύρεση μέγιστου y μεταξύ ιστογράμματος και πυκνότητας
> tmp.hist <- hist(rse, plot=FALSE, breaks="FD")
> tmp.hist     #  tmp.hist$densities πυκνότητα σχετικής συχνότητας
.........................
$density
 [1] 0.1600000 0.1000000 0.0850000 0.0700000 0.0300000 0.0200000
 [7] 0.0150000 0.0000000 0.0000000 0.0100000 0.0000000 0.0050000
[13] 0.0050000
.............................
> tmp.dens <- dexp(0,rate=1/5)   # λ*exp(-l*0) μέγιστο της πυκνότητας
> tmp.dens
[1] 0.2
> y.max <- max(tmp.hist$density,tmp.dens)
> # ΤΕΛΟΣ
> hist(rse, ylim=c(0,y.max), prob=TRUE, breaks="FD", col=gray(0.9),
main="Exp(1/5)")
> x1 <- seq(0,25,0.01)
> y1 <- 0.2*exp(-0.2*x1)
> lines(x1,y1)
> rug(rse)
```



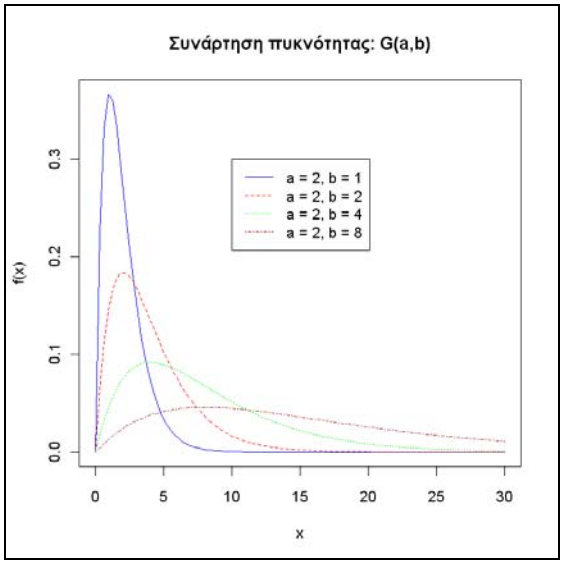
5.2.5 Κατανομή Γάμμα

Η κατανομή Γάμμα με παράμετρο μορφής a και κλίμακας b (συμβ. $G(a,b)$) έχει συνάρτηση πυκνότητας που δίνεται από τον τύπο

$$f(x)=\frac{1}{b^a\Gamma(a)}x^{a-1}e^{-x/b},\quad x\geq 0\,.$$

Με την εκτέλεση της εντολής "> `?GammaDist(stats)`" προκύπτουν πληροφορίες για την κατανομή Γάμμα. Για την κατανομή $G(2,1)$ δίνουμε την ακόλουθη γραφική παράσταση.

```
> x <- seq(0,30,length=100)
> plot(x,dgamma(x,shape=2,scale=1), type="l", col="blue", xlab="x",
+ ylab="f(x)", main="Συνάρτηση πυκνότητας: G(a,b)")
> lines(x,dgamma(x,shape=2,scale=2), col="red", lty=2)
> lines(x,dgamma(x,shape=2,scale=4), col="green", lty=3)
> lines(x,dgamma(x,shape=2,scale=8), col="brown", lty=4)
> legend(x=10,y=.3,paste("a = 2, b =", c(1,2,4,8)),lty=1:4,
+ col=c("blue","red","green","brown"))
```



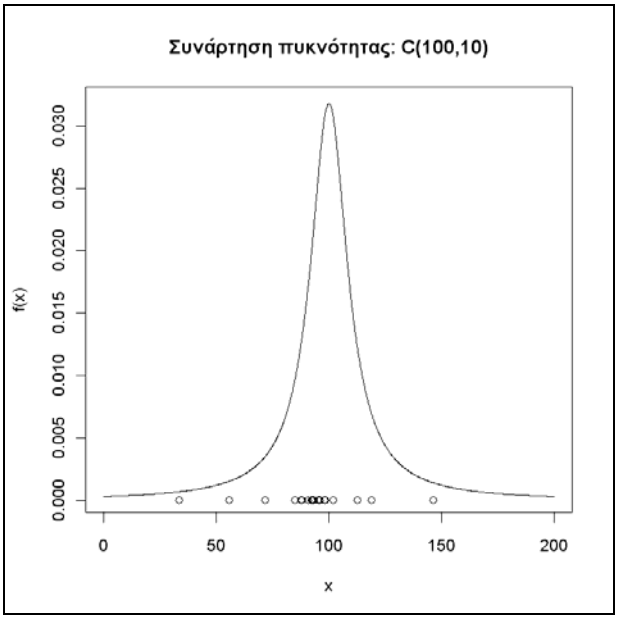
5.2.6 Κατανομή Cauchy

Η κατανομή Cauchy με παράμετρο θέσης a και κλίμακας b (συμβ. $C(a,b)$) έχει συνάρτηση πυκνότητας που δίνεται από τον τύπο

$$f(x)=\frac{1}{\pi b(1+[(x-a)/b]^2)},\quad -\infty<x<\infty.$$

Πληροφορίες για την κατανομή Cauchy προκύπτουν με την εκτέλεση της εντολής "> `?Cauchy(stats)`". Για την κατανομή $C(100,10)$ δίνουμε το ακόλουθο γράφημα.

```
> x <- seq(0,200,0.1)
> rsc <- rcauchy(20,100,10)
> plot(x,dcauchy(x,100,10), type="l", xlab="x", ylab="f(x)",
+ main="Συνάρτηση πυκνότητας: C(100,10)")
> points(rsc, rep(0,length(rsc)))
```



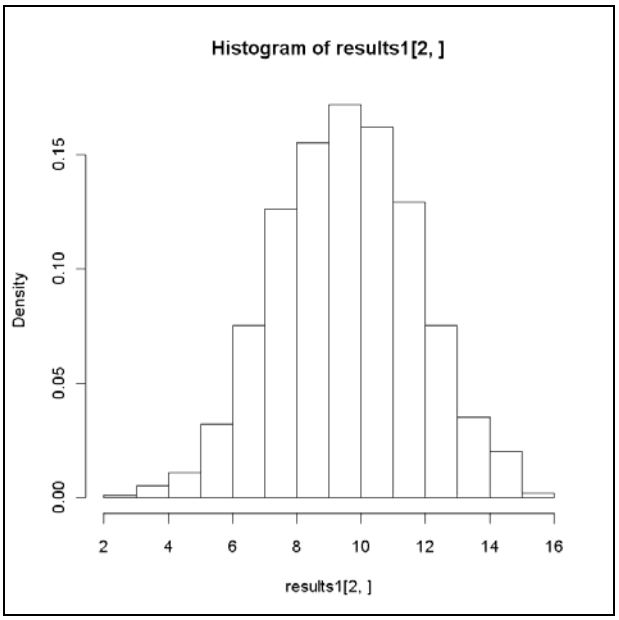
5.2.7 Πολυωνυμική κατανομή

Η τυχαία μεταβλητή $\mathbf{X}'=(X_1,X_2,...,X_k)$ ακολουθεί την πολυωνυμική κατανομή με παραμέτρους n, p_1, p_2, ..., p_k (συμβ. $M(n,p_1,p_2,...,p_k)$) αν η από κοινού συνάρτηση πιθανότητας των $X_1,X_2,...,X_k$ δίνεται από τον τύπο

$$f(x_1,x_2,...,x_k)=\frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k},\quad \sum p_i=1,\quad \sum x_i=n.$$

Πληροφορίες για την πολυωνυμική κατανομή προκύπτουν με την εκτέλεση της εντολής "> `?Multinomial(stats)`". Για την πολυωνυμική κατανομή $M(20,0.2,0.5,0.3)$ δίνουμε το ακόλουθο παράδειγμα

```
> n <-20; p <- c(0.2,0.5,0.3); x <- c(4,10,6)
> dmultinom(x, n, p)                        # f(4,10,6)
[1] 0.04419421
> factorial(n)/(factorial(x[1])*factorial(x[2])*factorial(x[3]))*
+ (p[1])^x[1]*(p[2])^x[2]*(p[3])^x[3]      # Επιβεβαίωση
[1] 0.04419421
> results <-rmultinom(1000, 20, c(0.2,0.5,0.3))  # Τυχαίο δείγμα μεγέθους 1000
> results
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,]    7    4    2    7    3    4    7    6    9     4     1     4     5     4
[2,]    9   11    7    8   14   10    6   10    9    10    10     7    12    12
[3,]    4    5   11    5    3    6    7    4    2     6     9     9     3     4

.............................................................

     [,999] [,1000]
[1,]      4       3
[2,]     10      10
[3,]      6       7
> mean(results1[2,])                  # Δειγματικός μέσος της X2
[1] 9.918
> hist(results[2,], prob=TRUE)        # Ιστόγραμμα της X2
```



5.2.8 Πολυδιάστατη κανονική κατανομή

Η δισδιάστατη τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2)$ ακολουθεί την δισδιάστατη κανονική κατανομή με παραμέτρους μ_1, μ_2, σ_1^2, σ_2^2, και ρ αν η από κοινού συνάρτηση πυκνότητας των X_1, X_2 δίνεται από τον τύπο

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}T^2\right), \quad x_1, x_2 \in R,$$

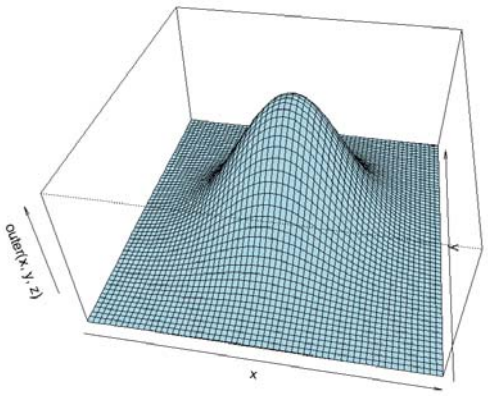
όπου

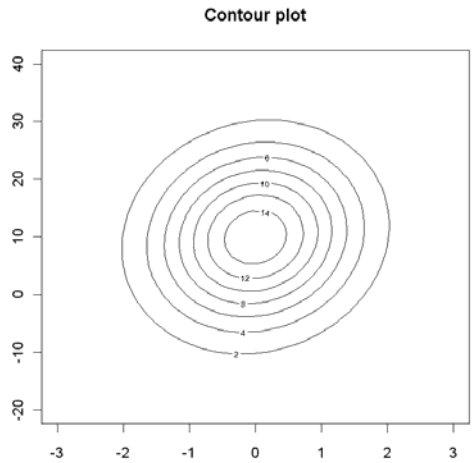
$$T^2 = \frac{1}{1-\rho^2}\cdot\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]$$

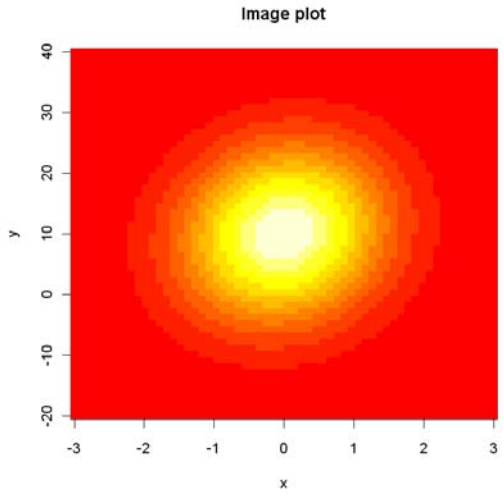
$$= \frac{1}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}\cdot[\sigma_2^2(x_1-\mu_1)^2+\sigma_1^2(x_2-\mu_2)^2-2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)].$$

Για τη διδιάστατη κανονική κατανομή δίνονται οι ακόλουθες γραφικές παραστάσεις (γίνεται χρήση των συναρτήσεων `function, persp, contour` και `image`)

```
> m1 <- 0; s1 <- 1; m2 <- 10; s2 <- 10; r <- 0.1
> x1 <- m1-3*s1; x2 <- m1+3*s1; y1 <- m2-3*s2; y2 <- m2+3*s2
> x <- seq(x1,x2,length=60)
> y <- seq(y1,y2,length=60)
> z <- function(x,y)(1/2*pi*s1*s2*(sqrt(1-r^2)))*exp(-0.5*(1/(1-r^2))*
+ ((((x-m1)/s1)^2)+(((y-m2)/s2)^2)-2*r*((x-m1)/s1)*((y-m2)/s2)))
> persp(x,y,outer(x,y,z),xlim = range(x), ylim = range(y),
+ theta = 10, phi = 40, expand = 0.5, col = "lightblue")
> contour(x,y, outer(x,y,z), nlevels=10, main="Contour plot")
> image(x,y, outer(x,y,z), main="Image plot")
```







Η τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2, ..., X_p)$ ακολουθεί την p-διάστατη κανονική κατανομή με παραμέτρους $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ (συμβολισμός $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), όπου

$$\boldsymbol{\mu}' = [\mu_1, \mu_2, ..., \mu_p]$$

είναι ένα διάνυσμα πραγματικών αριθμών και

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

είναι ένας συμμετρικός θετικά ορισμένος πίνακας, αν η συνάρτηση πυκνότητάς της δίνεται από τον τύπο

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left(-\frac{1}{2} T^2 \right), \quad \mathbf{x} \in R^p$$

(με $|\boldsymbol{\Sigma}|$ συμβολίζουμε την ορίζουσα του πίνακα $\boldsymbol{\Sigma}$ και με $\boldsymbol{\Sigma}^{-1}$ τον αντίστροφο του πίνακα $\boldsymbol{\Sigma}$). Στην ειδική περίπτωση $p = 1$ προκύπτει η συνήθης (μονοδιάστατη) κανονική κατανομή ενώ για $p = 2$ προκύπτει η δισδιάστατη κανονική κατανομή. Μπορεί να δειχθεί ότι το διάνυσμα $\boldsymbol{\mu}$ αποτελεί το διάνυσμα των μέσων τιμών και ο πίνακας $\boldsymbol{\Sigma}$ αποτελεί των πίνακα διακυμάνσεων-συνδιακυμάνσεων της τυχαίας μεταβλητής $\mathbf{X}$, δηλαδή

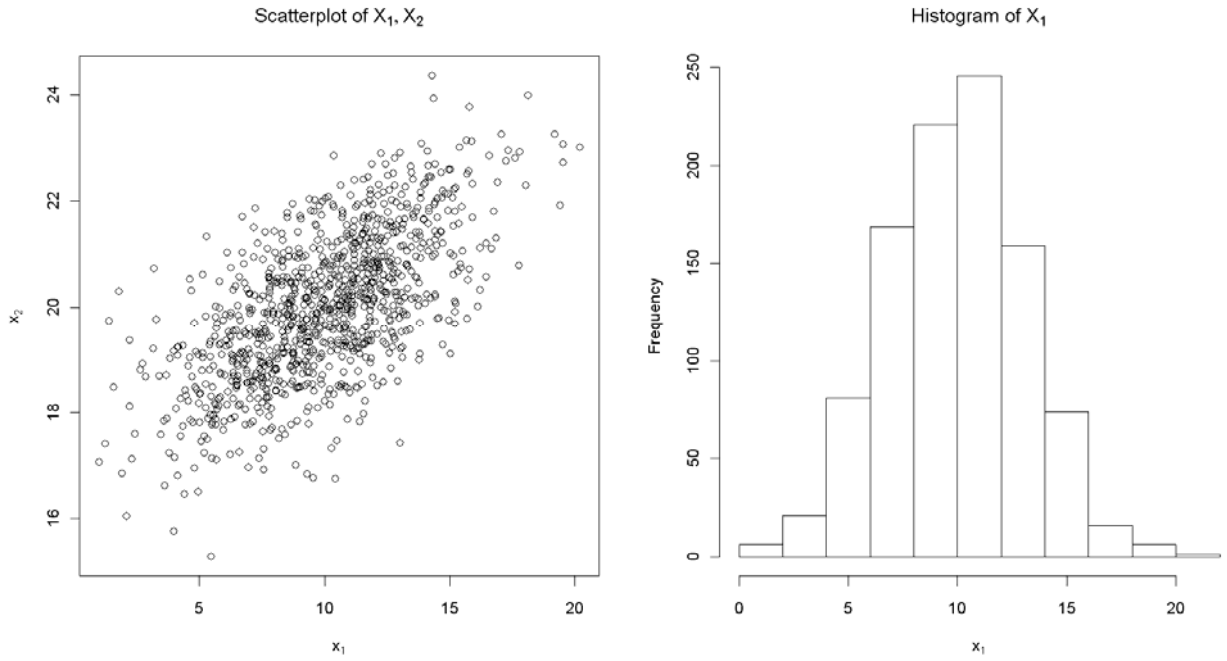
$$E(\mathbf{X}) = \boldsymbol{\mu}, \qquad E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}.$$

Για τη δημιουργία τυχαίων δειγμάτων από την $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ χρησιμοποιείται η συνάρτηση `mvrnorm` του πακέτου MASS. Η μορφή της συνάρτησης `mvrnorm` είναι η `mvrnorm(n, mu, Sigma)` όπου mu $= \boldsymbol{\mu}$ και Sigma $= \boldsymbol{\Sigma}$. Για τη δισδιάστατη περίπτωση δίνουμε το ακόλουθο παράδειγμα που χρησιμοποιεί ένα τυχαίο δείγμα μεγέθους 1000 από την κατανομή $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ με

$$\boldsymbol{\mu}' = [\mu_1, \mu_2] = [10, 20], \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 3 \\ 3 & 2 \end{bmatrix}.$$

```
> library(MASS)
> V <- matrix(c(10,3,3,2),2,2)
> V
     [,1] [,2]
[1,]   10    3
[2,]    3    2
> data <- mvrnorm(n=1000, mu=c(10,20), Sigma=V)
> a <- var(data); a

          [,1]     [,2]
[1,] 10.912195 3.311684
[2,]  3.311684 2.087406
```

```
> cor(data)
          [,1]     [,2]
[1,] 1.00000 0.69312
[2,] 0.69312 1.00000
> cor <- a[1,2]/((sqrt(a[1,1]))*(sqrt(a[2,2])))
> cor
[1] 0.69312
> cor(data[,1],data[,2])
[1] 0.69312
> par(mfrow=c(1,2))
> plot(data, xlab=expression(x[1]), ylab=expression(x[2]),
main=expression("Scatterplot of "*list(X[1],X[2])))
> x1 <- data[,1]
> hist(x1, xlab=expression(x[1]), main=expression("Histogram of "*X[1]))
```



5.3 Σύνοψη εντολών Κεφαλαίου 5

```
choose, contour, curve
dbinom, dgamma, dgeom, digits, dmultinom, dnorm dgamma
ecdf
factorial, fig, function
image
mvrnorm
options
pbinom, persp, plot.stepfun, pnorm, points, polygon
qnorm
reom, rexp, rmultinom, rnorm, rug
sample, set.seed
```

ΚΕΦΑΛΑΙΟ 6

Θέματα Στατιστικής με το R

6.1 Εισαγωγή

Στο παρόν κεφάλαιο θα ασχοληθούμε κυρίως με ελέγχους στατιστικών υποθέσεων, ενώ στην τελευταία παράγραφο θα αναπτύξουμε περιληπτικά βασικές έννοιες στην απλή και πολλαπλή γραμμική παλινδρόμηση.

Για την εκτέλεση ενός ελέγχου υπόθεσης για την παράμετρο θ ενός πληθυσμού X με τη βοήθεια ενός τυχαίου δείγματος $\mathbf{X} = (X_1, X_2, ..., X_n)$ από τον πληθυσμό, βασιζόμαστε στην παρατηρούμενη τιμή $u = u(\mathbf{x})$ μιας κατάλληλης στατιστικής συνάρτησης ελέγχου $U = U(\mathbf{X})$. Η κρίσιμη πληροφορία που δίνουν τα στατιστικά πακέτα για να ληφθεί η απόφαση της αποδοχής ή της απόρριψης της μηδενικής υπόθεσης είναι η $p-value$ του ελέγχου. Η $p-value$ του ελέγχου είναι η πιθανότητα να παρατηρήσουμε στην τύχη, κάτω από μια απλή μηδενική υπόθεση H_0, μια τιμή της στατιστικής συνάρτησης ελέγχου που είναι ίδια με αυτή που ήδη παρατηρήσαμε, ή ακόμα πιο ακραία από την H_0. Έτσι αν $p-value < a$, όπου a είναι το επίπεδο σημαντικότητας του ελέγχου, η μηδενική υπόθεση απορρίπτεται. Για τους ελέγχους υποθέσεων

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta < \theta_0,$$

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta > \theta_0,$$

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta \neq \theta_0,$$

που αφορούν την παράμετρο μιας συνεχούς τυχαίας μεταβλητής η $p-value$ του ελέγχου υπολογίζεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	***p – value***
$H_1 : \theta < \theta_0$	$P(U \le u \mid H_0)$
$H_1 : \theta > \theta_0$	$P(U \ge u \mid H_0)$
$H_1 : \theta \neq \theta_0$	$2 \cdot \min\{P(U \le u \mid H_0), P(U \ge u \mid H_0)\}$

Το R παρέχει ειδικές συναρτήσεις για τη διεξαγωγή ελέγχου υποθέσεων της μορφής `όνομα_συνάρτησης.test`. Το πιο συνηθισμένο όρισμα σε ένα έλεγχο υπόθεσης είναι το «`alternative=`» με τιμές `"two.sided"`, `"less"` και `"greater"`, το οποίο καθορίζει τη μορφή της εναλλακτικής υπόθεσης. Στην περίπτωση που ο έλεγχος αφορά μια παράμετρο ενός πληθυ-

σμού (ή ακόμη και διαφορές παραμέτρων ορισμένες φορές), υπάρχει η δυνατότητα εμφάνισης διαστημάτων εμπιστοσύνης της παραμέτρου δίνοντας απλά τιμή στο όρισμα «`conf.level=`» ίση με τον επιθυμητό συντελεστή εμπιστοσύνης $(1-a)$ του διαστήματος. Επίσης, συνήθως, το όρισμα «`correct=`» με τιμές `TRUE` ή `FALSE` δηλώνει αν θα χρησιμοποιηθεί κάποιο είδος διόρθωσης συνέχειας, ενώ το όρισμα «`exact=`» με τιμές `TRUE` ή `FALSE` δηλώνει αν θα υπολογιστεί η ακριβής ή η προσεγγιστική τιμή της $p-value$ του ελέγχου.

6.2 Συμπερασματολογία για ένα δείγμα

6.2.1 Έλεγχος της μέσης τιμής ενός πληθυσμού (t-test, z-test)

Έστω ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από πληθυσμό $N(\mu, \sigma^2)$ με άγνωστη διακύμανση. Για τον έλεγχο των υποθέσεων

$$H_0 : \mu = \mu_0 \quad - \quad H_1 : \mu < \mu_0 ,$$

$$H_0 : \mu = \mu_0 \quad - \quad H_1 : \mu > \mu_0 ,$$

$$H_0 : \mu = \mu_0 \quad - \quad H_1 : \mu \neq \mu_0 ,$$

χρησιμοποιούμε τη στατιστική συνάρτηση ελέγχου

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

η οποία ακολουθεί την κατανομή t_{n-1} όταν η H_0 είναι αληθής. Αν συμβολίσουμε με t την παρατηρούμενη τιμή της T, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$p-value$
$H_1 : \mu < \mu_0$	$\{t : t < -t_{n-1;\,a}\}$	$P(T \le t)$
$H_1 : \mu > \mu_0$	$\{t : t > t_{n-1;\,a}\}$	$P(T \ge t)$
$H_1 : \mu \neq \mu_0$	$\{t : t < -t_{n-1;\,a/2}\} \cup \{t : t > t_{n-1;\,a/2}\}$	$2P(T \ge \lvert t \rvert) = 2(1 - P(T \le \lvert t \rvert))$

Από το Κεντρικό Οριακό Θεώρημα προκύπτει ότι για μεγάλο μέγεθος δείγματος $(n \ge 30)$ η προαναφερθείσα στατιστική συνάρτηση ελέγχου μπορεί κάλλιστα να χρησιμοποιηθεί για τον έλεγχο της μέσης τιμής ενός πληθυσμού χωρίς καμία περιοριστική υπόθεση ως προς την κατανομή του (μπορεί να είναι ακόμη και διακριτή κατανομή). Πάντως, αν η κατανομή του πληθυσμού διαφέρει αρκετά από την κανονική κατανομή (κυρίως για μικρά, αλλά και για μεγάλα δείγματα) τότε είναι προτιμότερο να χρησιμοποιούνται άλλοι μη παραμετρικοί έλεγχοι όπως, π.χ., ο έλεγχος Wilcoxon Signed-Rank.

Στο R χρησιμοποιείται η συνάρτηση `t.test` για τον έλεγχο της μέσης τιμής ενός πληθυσμού με άγνωστη διακύμανση. Η βασική σύνταξη της συνάρτησης `t.test` για τον έλεγχο της μέσης τιμής ενός πληθυσμού είναι η ακόλουθη

```
t.test(x, mu=μ0)
x: το δείγμα σε μορφή διανύσματος
mu: η τιμή της μέση τιμής υπό τη μηδενική υπόθεση
```

Περισσότερες πληροφορίες για τα ορίσματα της συνάρτησης `t.test` προκύπτουν εκτελώντας την εντολή `?t.test`.

Παράδειγμα 6.1 (t-test)

Από την περιοχή Α μιας χώρας επιλέχθηκαν τυχαία 60 άνδρες και μετρήθηκε το ύψος τους (σε cm). Οι μετρήσεις δίνονται στο ακόλουθο πλαίσιο (αρχείο HEIGHTA.txt)

172.2	174.8	174.1	173.9	169.4	169.6	174.0	179.9	174.4	174.0
173.9	171.5	179.2	172.3	173.6	169.6	170.4	178.0	177.2	176.7
175.0	172.3	179.5	174.2	174.9	179.0	174.3	174.0	177.0	172.0
174.9	178.1	170.8	174.1	172.6	174.0	171.5	178.4	179.0	171.3
174.8	176.1	175.0	176.2	175.3	174.0	177.4	175.1	175.1	175.9
171.4	166.7	177.3	179.2	178.2	176.2	172.0	173.6	182.1	172.9

Ο έλεγχος

$$H_0 : \mu_A = 174 \quad - \quad H_1 : \mu_A \neq 174$$

εκτελείται με το R ως εξής:

```
> HA <- read.table("HEIGHTA.txt", header=T)
> attach(HA)
> names(HA)
[1] "HEIGHTA"
> t.test(HEIGHTA, mu=174, alternative="two.sided", conf.level=0.95)

        One Sample t-test

data:  HEIGHTA
t = 1.7156, df = 59, p-value = 0.09148
alternative hypothesis: true mean is not equal to 174
95 percent confidence interval:
 173.8888 175.4478
sample estimates:
mean of x
 174.6683
```

Πέραν της $p-value$ του ελέγχου δίνεται και το 95% διάστημα εμπιστοσύνης για το μέσο του πληθυσμού (λόγω του ορίσματος `conf.level=0.95`).

Επιβεβαίωση της *p-value*:

```
> m <- mean(HEIGHTA); s <- sd(HEIGHTA); n <- length(HEIGHTA)
> t <- (m-174)/(s/sqrt(n))
> p1 <- pt(abs(t), df=n-1, lower.tail = FALSE)
> p2 <- pt(abs(t), df=n-1)
> cat("Επιβεβαίωση: p-value =", 2*min(p1,p2),"\n")
```

```
Επιβεβαίωση: p-value = 0.09147797
```

Στην περίπτωση που η διακύμανση του πληθυσμού είναι γνωστή (φυσικά κάτι τέτοιο ισχύει σπάνια στην πράξη) τότε χρησιμοποιείται η στατιστική συνάρτηση ελέγχου

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

η οποία ακολουθεί την τυποποιημένη (ή τυπική) κανονική κατανομή $N(0,1)$ όταν η H_0 είναι αληθής. Αν συμβολίσουμε με z την παρατηρούμενη τιμή της Z, η κρίσιμη περιοχή σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : \mu < \mu_0$	$\{z : z < -z_a\}$	$P(Z \le z) = \Phi(z)$
$H_1 : \mu > \mu_0$	$\{z : z > z_a\}$	$P(Z \ge z) 1 - \Phi(z)$
$H_1 : \mu \ne \mu_0$	$\{z : z < -z_{a/2}\} \bigcup \{z : z > z_{a/2}\}$	$2P(Z \ge \lvert z \rvert) = 2(1 - \Phi(\lvert z \rvert))$

Ο παραπάνω έλεγχος γίνεται στο R με τη συνάρτηση `z.test` του πακέτου BSDA (επίσης και με το πακέτο PASWR). Η βασική σύνταξη της συνάρτησης `z.test` είναι η ακόλουθη

```
z.test(x, mu=μ0, sigma.x=σ)
x: το δείγμα σε μορφή διανύσματος
mu: η τιμή της μέση τιμής υπό τη μηδενική υπόθεση
sigma.x: η τιμή της τυπικής απόκλισης του πληθυσμού
```

Παράδειγμα 6.2 (z-test)

Συνεχίζοντας το Παράδειγμα 6.1 ας υποθέτουμε ότι η τυπική απόκλιση του πληθυσμού του ύψους των ανδρών είναι $\sigma = 3$. Τότε ο έλεγχος

$$H_0 : \mu_A = 174 \quad - \quad H_1 : \mu_A \ne 174$$

εκτελείται με το R ως εξής:

```
> library(BSDA); s <- 3
> z.test(HEIGHTA, mu=174, sigma.x=s, alternative="two.sided",
conf.level=0.95)

        One-sample z-Test

data:  HEIGHTA
z = 1.7256, p-value = 0.08441
alternative hypothesis: true mean is not equal to 174
95 percent confidence interval:
 173.9092 175.4274
sample estimates:
mean of x
 174.6683
```

Επιβεβαίωση της *p-value*:

```
> z <- (m-174)/(s/sqrt(n))
> pvalue <- 2*(1-pnorm(abs(z)))
> cat("Επιβεβαίωση: p-value =", pvalue)
Επιβεβαίωση: p-value = 0.08441413
```

6.2.2 Έλεγχος της διακύμανσης ενός κανονικού πληθυσμού

Έστω ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από κανονικό πληθυσμό. Για τους ελέγχους

$$H_0 : \sigma^2 = \sigma_0^2 \quad - \quad H_1 : \sigma^2 < \sigma_0^2,$$

$$H_0 : \sigma^2 = \sigma_0^2 \quad - \quad H_1 : \sigma^2 > \sigma_0^2,$$

$$H_0 : \sigma^2 = \sigma_0^2 \quad - \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

βασιζόμαστε στη στατιστική συνάρτηση ελέγχου

$$U = \frac{(n-1)S^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma_0^2}$$

η οποία ακολουθεί την κατανομή χ^2_{n-1} όταν η H_0 είναι αληθής. Αν συμβολίσουμε με u την παρατηρούμενη τιμή της U, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στον ακόλουθο πίνακα

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : \sigma^2 < \sigma_0^2$	$\{u : u < \chi^2_{n-1;1-a}\}$	$P(U \le u)$
$H_1 : \sigma^2 > \sigma_0^2$	$\{u : u > \chi^2_{n-1;a}\}$	$P(U \ge u)$
$H_1 : \sigma^2 \neq \sigma_0^2$	$\{u : u < \chi^2_{n-1;1-a/2}\} \bigcup \{u : u > \chi^2_{n-1;a/2}\}$	$2\min\{P(U \le u), P(U \ge u)\}$

Παράδειγμα 6.3

Μια μηχανή γεμίζει σάκους των 25Kg με τσιμέντο. Υπό φυσιολογικές συνθήκες η ποσότητα του τσιμέντου που περιέχει ένας σάκος ακολουθεί την κανονική κατανομή με μέση τιμή 25Kg και διακύμανση 0.35Kg. Από την παραγωγή της συγκεκριμένης μηχανής επιλέχτηκαν 30 σάκοι (ένας κάθε ώρα) με βάρη που δίνονται το ακόλουθο πλαίσιο (αρχείο cement.txt)

i	X_i	i	X_i	i	X_i	i	X_i	i	X_i	i	X_i
1	26.18	6	25.44	11	24.22	16	26.24	21	24.22	26	25.84
2	25.30	7	24.49	12	26.48	17	25.46	22	24.49	27	26.09
3	25.18	8	25.01	13	23.97	18	25.01	23	25.68	28	25.21
4	24.54	9	25.12	14	25.83	19	24.71	24	26.01	29	26.04
5	25.14	10	25.67	15	25.05	20	25.27	25	25.50	30	25.23

Ο έλεγχος

$$H_0 : \sigma^2 = 0.35 \quad - \quad H_1 : \sigma^2 > 0.35$$

εκτελείται με το R ως εξής:

```
> cement <- read.table("cement.txt", header=T)
> names(cement); attach(cement)
[1] "x"
> v <- 0.35; u <- (length(x)-1)*var(x)/v
> pvalue <- 1-pchisq(u, length(x)-1)
> cat("p-value=", pvalue, "\n")
p-value= 0.1983238
```

Έτσι, για παράδειγμα, σε επίπεδο σημαντικότητας $a = 0.1$ η μηδενική υπόθεση δεν απορρίπτεται.

6.2.3 Έλεγχος της αναλογίας επιτυχιών σε ένα πληθυσμό Bernoulli

Έστω ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ από πληθυσμό $B(1, p)$. Για τον έλεγχο των υποθέσεων

$$H_0 : p = p_0 \quad - \quad H_1 : p < p_0,$$

$$H_0 : p = p_0 \quad - \quad H_1 : p > p_0,$$

$$H_0 : p = p_0 \quad - \quad H_1 : p \neq p_0,$$

βασιζόμαστε στη στατιστική συνάρτηση ελέγχου $X = \sum_{i=1}^{n} X_i$, η οποία ακολουθεί κατανομή $B(n, p_0)$ όταν η H_0 είναι αληθής. Αν συμβολίσουμε με x την παρατηρούμενη τιμή της X, τότε ο υπολογισμός της $p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	***p – value***
$H_1 : p < p_0$	$P(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
$H_1 : p > p_0$	$P(X \ge x) = \sum_{i=x}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$
$H_1 : p \neq p_0$	$\sum_{i=0}^{n} I[P(X=i) \le P(X=x)] \binom{n}{i} p_0^i (1-p_0)^{n-i}$

Σημειώνουμε ότι $I(A) = 1$ όταν η δήλωση A είναι αληθής, ειδάλλως $I(A) = 0$. Ο έλεγχος υπόθεσης για την αναλογία p των επιτυχιών σε ένα πληθυσμό Bernoulli γίνεται στο *R* με τη συνάρτηση `binom.test`. Η βασική σύνταξη της συνάρτησης `binom.test` για τον έλεγχο της αναλογίας επιτυχιών σε ένα πληθυσμό Bernoulli είναι η ακόλουθη

```
binom.test(x, n, p=p0)
x: αριθμός επιτυχιών (x= Σxi)
n: αριθμός δοκιμών (μέγεθος δείγματος)
p: η πιθανότητα επιτυχίας υπό τη μηδενική υπόθεση
```

Ο κώδικας για τη συνάρτηση `binom.test` εμφανίζεται εκτελώντας την εντολή `binom.test`, από τον οποίο ο έμπειρος χρήστης μπορεί να αντλήσει πολύτιμες πληροφορίες. Την ίδια τακτική μπορείτε να ακολουθήσετε και για τις υπόλοιπες συναρτήσεις του παρόντος κεφαλαίου.

Παράδειγμα 6.4 (ακριβής έλεγχος)

Σε ένα τυχαίο δείγμα μεγέθους 600 ατόμων βρέθηκαν 276 καπνιστές (επιτυχίες). Ο έλεγχος της υπόθεσης

$$H_0 : p = 0.4 \quad - \quad H_1 : p \neq 0.4$$

που αφορά την αναλογία (ποσοστό) p των καπνιστών στον πληθυσμό εκτελείται στο R ως εξής:

```
> n <- 600; x <- 276; p0=0.4
> binom.test(x, n, p0, conf.level=0.9)

        Exact binomial test

data:  x and n
number of successes = 276, number of trials = 600, p-value = 0.003063
alternative hypothesis: true probability of success is not equal to 0.4
90 percent confidence interval:
 0.4258999 0.4943934
sample estimates:
probability of success
                  0.46
```

Πέραν της $p-value$ του ελέγχου δίνεται εκτίμηση για την αναλογία p των επιτυχιών $(276/600 = 0.46)$ καθώς επίσης και το 90% διάστημα εμπιστοσύνης που δίνεται επειδή δηλώσαμε το όρισμα `conf.level=0.9` (για περισσότερες λεπτομέρειες σχετικά με τον τρόπο υπολογισμού του διαστήματος εμπιστοσύνης για την αναλογία p των επιτυχιών ο αναγνώστης παραπέμπεται στις αναφορές που παραθέτει το R στο help της συνάρτησης `binom.test` (`?binom.test`)).

Επιβεβαίωση της *p-value*:

```
> probs <- dbinom(0:n, n, p0)
> pvalue <- sum(probs[probs<=dbinom(x,n,p0)])
> cat("Επιβεβαίωση: p-value=", pvalue, "\n")
Επιβεβαίωση: p-value= 0.003063216
```

Οι παραπάνω έλεγχοι, στην περίπτωση που το μέγεθος του δείγματος είναι μεγάλο, μπορούν να γίνουν (προσεγγιστικά) χρησιμοποιώντας την προσέγγιση της διωνυμικής κατανομής από την κανονική κατανομή. Η στατιστική συνάρτηση ελέγχου είναι η

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np_0}{\sqrt{p_0(1-p_0)}} = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

η οποία, όταν η H_0 είναι αληθής, για μεγάλα n ακολουθεί προσεγγιστικά την κατανομή $N(0,1)$. Αν συμβολίσουμε με z την παρατηρούμενη τιμή της Z, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : p < p_0$	$\{z : z < -z_a\}$	$P(Z \le z) = \Phi(z)$
$H_1 : p > p_0$	$\{z : z > z_a\}$	$P(Z \ge z) = 1 - \Phi(z)$
$H_1 : p \ne p_0$	$\{z : z < -z_{a/2}\} \cup \{z : z > z_{a/2}\}$	$2P(Z \ge \lvert z \rvert) = 2(1 - \Phi \lvert z \rvert)$

Οι παραπάνω έλεγχοι γίνονται στο R με τη συνάρτηση `prop.test`. Η βασική σύνταξη της συνάρτησης `prop.test` για τον έλεγχο της αναλογίας επιτυχιών σε ένα πληθυσμό Bernoulli είναι η ακόλουθη

```
prop.test(x, n, p=p0)
x: αριθμός επιτυχιών (x= Σxi)
n: αριθμός δοκιμών (μέγεθος δείγματος)
p: η πιθανότητα επιτυχίας υπό τη μηδενική υπόθεση
```

Παράδειγμα 6.5 (προσεγγιστικός έλεγχος)

Εφαρμόζοντας προσεγγιστικό έλεγχο στα δεδομένα του Παραδείγματος 6.5 παίρνουμε τα ακόλουθα αποτελέσματα:

```
> n <- 600; x <- 276; p0=0.4
> prop.test(x, n, p0, alternative="two.sided", correct=FALSE)

        1-sample proportions test without continuity correction

data:  x out of n, null probability p0
X-squared = 9, df = 1, p-value = 0.0027
alternative hypothesis: true p is not equal to 0.4
```

Παρατηρούμε ότι το R δίνει ως τιμή της στατιστικής συνάρτησης ελέγχου την z^2 (προφανώς η Z^2, όταν η μηδενική υπόθεση είναι αληθής, ακολουθεί την κατανομή χ^2_1.

Επιβεβαίωση της *p-value*:

```
> z <- ((x/n)-p0)/sqrt(p0*(1-p0)/n)
> pvalue <- 2*(1-pnorm(abs(z)))
> cat("Επιβεβαίωση: p-value=", pvalue, ", z=", z ,"\n")
Επιβεβαίωση: p-value= 0.002699796 , z= 3
```

Όταν $|(x/n) - p_0| > 1/2n$ (ή ισοδύναμα $|x - np_0| > 1/2$) χρησιμοποιείται συνήθως διόρθωση συνέχειας στην παρατηρούμενη τιμή z της Z, η οποία δίνεται στον ακόλουθο πίνακα

Συνθήκη	z	Συνθήκη	z
$(x/n)-p_0>0$	$\dfrac{(x/n)-p_0-(1/2n)}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$(x/n)-p_0<0$	$\dfrac{(x/n)-p_0+(1/2n)}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Ωστόσο η χρησιμοποίηση διόρθωσης συνέχειας οδηγεί σε πιο συντηρητικό έλεγχο (conservative test) αφού είναι πλέον λιγότερο πιθανό να απορριφθεί εσφαλμένα μια αληθή μηδενική υπόθεση και επομένως έχουμε μικρότερη ισχύ (η *p-value* του ελέγχου είναι μεγαλύτερη σε σχέση με τον έλεγχο που δεν χρησιμοποιεί διόρθωση συνέχειας).

Παράδειγμα 6.6 (προσεγγιστικός έλεγχος με διόρθωση συνέχειας)

Εφαρμόζοντας προσεγγιστικό έλεγχο με διόρθωση συνέχειας στα δεδομένα του Παραδείγματος 6.4 παίρνουμε τα ακόλουθα αποτελέσματα (ισχύει η συνθήκη $|(x/n)-p_0|>1/2n$).

```
> n <- 600; x <- 276; p0=0.4
> prop.test(x, n, p0, alternative="two.sided", correct=TRUE)

        1-sample proportions test with continuity correction

data:  x out of n, null probability p0
X-squared = 8.7517, df = 1, p-value = 0.003093
alternative hypothesis: true p is not equal to 0.4
```

Επιβεβαίωση της *p-value*:

```
> abs((x/n)-p0)>1/(2*n)
[1] TRUE   ## Χρειάζεται διόρθωση συνέχειας
> (x/n)>p0
[1] TRUE  ## Με -1/2n αντί +(1/2n
> z <- ((x/n)-p0-1/(2*n))/sqrt(p0*(1-p0)/n)
> pvalue <- 2*(1-pnorm(abs(z)))
> cat("Επιβεβαίωση: p-value =", pvalue, ", z =", z,"\n")
Επιβεβαίωση: p-value = 0.003093075 , z = 2.958333
```

6.2.4 Προσημικός έλεγχος (sign test)

Έστω $X_1, X_2, ..., X_n$ τυχαίο δείγμα από συνεχή πληθυσμό X με άγνωστη διάμεσο M. Μας ενδιαφέρουν οι έλεγχοι

$$H_0: M = M_0 \quad - \quad H_1: M < M_0,$$

$$H_0: M = M_0 \quad - \quad H_1: M > M_0,$$

$$H_0: M = M_0 \quad - \quad H_1: M \neq M_0.$$

Μια ισοδύναμη μορφή της μηδενικής υπόθεσης είναι η

$$H_0: P(X > M_0) = P(X < M_0) = 0.5.$$

Η στατιστική συνάρτηση ελέγχου S ορίζεται ως ο συνολικός αριθμός των θετικών διαφορών

$X_1 - M_0$, $X_2 - M_0$, ..., $X_n - M_0$. Προφανώς η S ακολουθεί κατανομή $B(n, 0.5)$ όταν η μηδενική υπόθεση είναι αληθής. Μεγάλες τιμές της S ($S >> n/2$) υποδηλώνουν ότι $M > M_0$. Αν υπάρχουν μηδενικές διαφορές τότε συνήθως τις παραλείπουμε και μειώνεται ισόποσα το μέγεθος του δείγματος n. Αν συμβολίσουμε με s την παρατηρούμενη τιμή της *S*, τότε ο υπολογισμός της $p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	$p-value$
$H_1 : M < M_0$	$P(S \le s) = \sum_{i=0}^{s} \binom{n}{i} 2^{-n}$
$H_1 : M > M_0$	$P(S \ge s) = \sum_{i=s}^{n} \binom{n}{i} 2^{-n}$
$H_1 : M \ne M_0$	$\sum_{i=0}^{n} I[P(S=i) \le P(S=s)] \binom{n}{i} 2^{-n}$

Αξίζει να σημειώσουμε ότι

$$\sum_{i=0}^{n} I[P(S=i) \le P(S=s)] \binom{n}{i} 2^{-n} = 2 \cdot \sum_{i=0}^{s^*} \binom{n}{i} 2^{-n}, \quad s^* = \min\{s, n-s\}.$$

Εναλλακτικά, οι παραπάνω έλεγχοι στην περίπτωση που το μέγεθος του δείγματος είναι σχετικά μεγάλο ($n > 20$) μπορούν να εκτελεστούν με την κανονική προσέγγιση της διωνυμικής κατανομής. Η στατιστική συνάρτηση ελέγχου τότε είναι η

$$Z = \frac{S - E(S)}{\sqrt{Var(S)}} = \frac{S - n/2}{\sqrt{n}/2}$$

Αν συμβολίσουμε με z την παρατηρούμενη τιμή της *Z*, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : M < M_0$	$\{z : z < -z_a\}$	$P(Z \le z)$
$H_1 : M > M_0$	$\{z : z > z_a\}$	$P(Z \ge z)$
$H_1 : M \ne M_0$	$\{z : z < -z_{a/2}\} \cup \{z : z > z_{a/2}\}$	$2P(Z \ge \lvert z \rvert) = 2(1 - P(Z \le \lvert z \rvert))$

Ο έλεγχος προσήμου μπορεί να εφαρμοστεί γενικότερα σε διατάξιμες παρατηρήσεις (ordinal data). Ένα "μειονέκτημα" του προσημικού ελέγχου είναι ότι δεν λαμβάνει υπόψη το μέγεθος των διαφορών $X_i - M_0$, παρά μόνο πόσες από αυτές είναι θετικές. Ο έλεγχος της επόμενης παραγράφου (Wilcoxon signed-rank) λαμβάνει υπόψη το μέγεθος των διαφορών αντιστοιχώντας μεγάλο βαθμό (rank) στις μεγαλύτερες διαφορές, και μικρό στις μικρότερες.

Ο προσημικός έλεγχος γίνεται στο R με τη συνάρτηση `SIGN.test` του πακέτου BSDA (επίσης και με το πακέτο PASWR). Η βασική σύνταξη της συνάρτησης `SIGN.test` είναι η ακόλουθη

```
SIGN.test(x, md=M0)
x: το δείγμα σε μορφή διανύσματος
md: η τιμή της διαμέσου υπό τη μηδενική υπόθεση
```

Παράδειγμα 6.7 (sign.test)

Για την παραγωγή του στιγμιαίου καφέ χρησιμοποιούνται δύο μέθοδοι: ψυχρό στέγνωμα και στέγνωμα μετά από ψεκασμό. Η διάμεσος του ποσού της καφεΐνης που μένει ως υπόλοιπο με τη μέθοδο του ψυχρού στεγνώματος είναι 3.5 γραμμάρια ανά 100 γραμμάρια ξηρού προϊόντος. Ακολούθως δίνεται το υπόλοιπο της καφεΐνης (σε γραμμάρια) με τη μέθοδο του στεγνώματος μετά από ψεκασμό που βρέθηκε σε 8 διαφορετικές μάρκες καφέ (ανά 100 γραμμάρια προϊόντος)

4.8, 4, 3.8, 4.3, 3.9, 4.6, 3.1, 3.7.

Για τον έλεγχο

$$H_0 : M = 3.5 \quad - \quad H_1 : M \neq 3.5$$

χρησιμοποιώντας το R παίρνουμε:

```
> library(BSDA)
> x <- c(4.8, 4, 3.8, 4.3, 3.9, 4.6, 3.1, 3.7)
> SIGN.test(x, md=3.5, alternative="t")

        One-sample Sign-Test

data:  x
s = 7, p-value = 0.07031
alternative hypothesis: true median is not equal to 3.5
95 percent confidence interval:
 3.505 4.665
sample estimates:
median of x
      3.95

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9297  3.700  4.600
Interpolated CI       0.9500  3.505  4.665
Upper Achieved CI     0.9922  3.100  4.800
```

Επιβεβαίωση της *p-value*:

```
> pvalue <- dbinom(0,8,0.5)+dbinom(1,8,0.5)+
+ dbinom(7,8,0.5)+dbinom(8,8,0.5)
> cat("Επιβεβαίωση: p-value=", pvalue, "\n")
Επιβεβαίωση: p-value= 0.0703125
```

6.2.5 Έλεγχος Wilcoxon Signed-Rank

Έστω $X_1, X_2, ..., X_n$ τυχαίο δείγμα από συνεχή πληθυσμό X ο οποίος είναι συμμετρικός γύρω από το σημείο θ. Μας ενδιαφέρουν οι έλεγχοι

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta < \theta_0 ,$$

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta > \theta_0 ,$$

$$H_0 : \theta = \theta_0 \quad - \quad H_1 : \theta \neq \theta_0 .$$

Οι παραπάνω έλεγχοι μπορούν να θεωρηθούν ότι είναι έλεγχοι που αφορούν τη μέση τιμή ή τη διάμεσο του πληθυσμού αφού έχει γίνει η υπόθεση ότι η κατανομή του πληθυσμού είναι συμμετρική.

Για την εκτέλεση των ελέγχων σχηματίζουμε αρχικά τις διαφορές $D_i = X_i - \theta_0$, υπολογίζουμε τις απόλυτες τιμές τους $|D_i|$ και αντιστοιχούμε βαθμούς (ranks) R_i στις διατεταγμένες τιμές $|D_i|$ (ξεκινάμε από τη μικρότερη απόλυτη τιμή $|D_i|$ στην οποία αντιστοιχούμε το βαθμό 1, στη δεύτερη μικρότερη αντιστοιχούμε το βαθμό 2, κ.ο.κ.). Στη συνέχεια υπολογίζουμε τις στατιστικές συναρτήσεις ελέγχου T^+ και T^- που δίνονται από τους τύπους

$$T^+ = \sum_{i=1}^{n} k_i R_i , \qquad T^- = \sum_{i=1}^{n} (1 - k_i) R_i$$

όπου

$$k_i = \begin{cases} 1, & \text{αν } D_i > 0 \\ 0, & \text{αν } D_i < 0. \end{cases}$$

Προφανώς

$$T^+ + T^- = \sum_{i=1}^{n} R_i = \frac{n(n+1)}{2}$$

και επομένως αν γνωρίζουμε την τιμή μιας εκ των δύο στατιστικών συναρτήσεων, τότε ο υπολογισμός της άλλης τιμής είναι άμεσος.

Η στατιστική συνάρτηση ελέγχου T^+ αθροίζει τους βαθμούς R_i που αντιστοιχούν στις παρατηρήσεις που υπερβαίνουν το θ_0, ενώ η στατιστική συνάρτηση ελέγχου T^- αθροίζει τους βαθμούς που αντιστοιχούν στις παρατηρήσεις που είναι μικρότερες από το θ_0. Συνεπώς αν τα T^+ και T^- διαφέρουν αρκετά τότε υπάρχουν ενδείξεις ότι $\theta \neq \theta_0$, αν $T^+ > T^-$ (ή ισοδύναμα για μεγάλες τιμές του T^+) τότε υπάρχουν ενδείξεις ότι $\theta > \theta_0$, ενώ αν $T^+ < T^-$ (ή ισοδύναμα για μικρές τιμές του T^+) τότε υπάρχουν ενδείξεις ότι $\theta < \theta_0$.

Αν συμβολίσουμε με t^+ την παρατηρούμενη τιμή της T^+ τότε ο (ακριβής) υπολογισμός της

$p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	$\boldsymbol{p-value}$
$H_1: \theta < \theta_0$	$P(T^+ \le t^+)$
$H_1: \theta > \theta_0$	$P(T^+ \ge t^+)$
$H_1: \theta \ne \theta_0$	$2\min\{P(T^+ \le t^+), P(T^+ \ge t^+)\}$

Η κατανομή της T^+ είναι διακριτή και παίρνει τιμές από 0 έως $n(n+1)/2$ (την τιμή 0 την παίρνει όταν όλες οι διαφορές $D_i = X_i - \theta_0$ είναι αρνητικές, ενώ την τιμή $n(n+1)/2$ όταν όλες οι διαφορές είναι θετικές). Μπορεί να δειχθεί ότι όταν η H_0 είναι αληθής, τότε η κατανομή της T^+ είναι συμμετρική με

$$E(T^+) = \frac{n(n+1)}{4}, \quad Var(T^+) = \frac{n(n+1)(2n+1)}{24}.$$

Επομένως, λόγω συμμετρικότητας της T^+ και επειδή $T^+ + T^- = n(n+1)/2$, η T^+ και η T^- έχουν την ίδια κατανομή. Πίνακες που περιλαμβάνουν τιμές της συνάρτησης κατανομής της T^+ (ή T^-) μπορούν να βρεθούν σε βιβλία μη παραμετρικής στατιστικής και επομένως μπορεί να γίνει ακριβής υπολογισμός της $p-value$ του ελέγχου.

Αποδεικνύεται ότι κάτω από την H_0 η στατιστική συνάρτηση

$$T^* = \frac{T^+ - E(T^+)}{\sqrt{Var(T^+)}} = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

ακολουθεί ασυμπτωτικά ($n \to \infty$) την κατανομή $N(0,1)$. Αν συμβολίσουμε με t^* την παρατηρούμενη τιμή της T^*, τότε ο προσεγγιστικός υπολογισμός της $p-value$ του ελέγχου (ο οποίος προτείνεται να χρησιμοποιείται για $n \ge 20$) γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	$\boldsymbol{p-value}$
$H_1: \theta < \theta_0$	$P(T^* \le t^*) = \Phi(t^*)$
$H_1: \theta > \theta_0$	$P(T^* \ge t^*) = 1 - \Phi(t^*)$
$H_1: \theta \ne \theta_0$	$2P(T^* \ge \lvert t^* \rvert) = 2(1 - \Phi(\lvert t^* \rvert))$

Σημειώνεται ότι αν κάποιες από τις διαφορές D_i είναι ίσες με το μηδέν, τότε τις αποβάλλουμε και μειώνουμε ανάλογα το μέγεθος του δείγματος. Επίσης αν υπάρχουν δεσμοί (ισοπαλίες, ties) στα $|D_i|$, τότε χρησιμοποιείται η στατιστική συνάρτηση

$$T^* = \frac{T^+ - E(T^+)}{\sqrt{Var(T^+)}} = \frac{T^+ - n(n+1)/4}{\sqrt{\frac{1}{24}\left(n(n+1)(2n+1) - \frac{1}{2}\sum_{j=1}^{r}(t_j^3 - t_j)\right)}}$$

η οποία ακολουθεί ασυμπτωτικά την κατανομή $N(0,1)$ όταν η H_0 είναι αληθής. Στον παραπάνω τύπο το r δηλώνει το πλήθος των διαφορετικών βαθμών, και το t_j δηλώνει πόσες φορές εμφανίζεται κάθε διαφορετικός βαθμός $(1 \le j \le r)$. Παρατηρούμε ότι η παρουσία των δεσμών δεν επηρεάζει τη μέση τιμή της T^+, επηρεάζει όμως τη διακύμανσή της.

Ο έλεγχος Wilcoxon Signed-Rank ισχύει γενικότερα για παρατηρήσεις $X_1, X_2, ..., X_n$ οι οποίες δεν είναι απαραίτητο να προέρχονται από τον ίδιο πληθυσμό (αρκεί φυσικά να ικανοποιούνται οι υπόλοιπες δύο προϋποθέσεις, δηλαδή οι παρατηρήσεις να είναι ανεξάρτητες και να προέρχονται από συμμετρικούς πληθυσμούς γύρω από το ίδιο σημείο θ). Επίσης, στην πράξη, δεν υπάρχει μεγάλη εμμονή γύρω από την υπόθεση του να προέρχονται οι παρατηρήσεις από συνεχή πληθυσμό, αρκεί να είναι διατάξιμα (ordinal data). Τέλος, αξίζει να σημειωθεί ότι ο έλεγχος Wilcoxon Signed-Rank προτείνεται να χρησιμοποιείται για τον έλεγχο της μέσης τιμής ενός πληθυσμού αντί του `t.test` όταν η κατανομή του πληθυσμού απέχει αρκετά από την κανονική κατανομή.

Στο R ο έλεγχος Wilcoxon Signed-Rank για ένα δείγμα πραγματοποιείται με τη συνάρτηση `wilcox.test`. Η βασική σύνταξη της συνάρτησης `wilcox.test` είναι η ακόλουθη

```
wilcox.test(x, mu=θ0, exact=TRUE or FALSE, correct=TRUE or FALSE)
x: το δείγμα σε μορφή διανύσματος
mu: η τιμή του θ (διάμεσος/μέση τιμή) υπό τη μηδενική υπόθεση
exact: ακριβής (TRUE) ή προσεγγιστικός (FALSE) υπολογισμός της
    p.value (αν δεν δηλωθεί το όρισμα exact τότε για n<50 παίρνου-
    με ακριβή p-value)
correct: υπολογισμός p-value με διόρθωση (TRUE) ή χωρίς διόρθωση
    (FALSE) συνέχειας (αν δεν δηλωθεί το όρισμα correct έχει de-
    fault τιμή TRUE)
```

Παράδειγμα 6.8 (προσεγγιστικός έλεγχος)

Ένας γιατρός ισχυρίζεται ότι η διάμεσος του αριθμού των φορών που βλέπει κάθε ασθενή του στη διάρκεια ενός έτους είναι 5. Σε ένα τυχαίο δείγμα μεγέθους 13 ασθενών του βρήκε ότι ο αριθμός των επισκέψεών τους στη διάρκεια του προηγούμενου έτους ήταν

5, 9, 10, 8, 4, 8, 5, 3, 0, 10, 15, 9, 5

αντίστοιχα. Θέλουμε να εκτελέσουμε το έλεγχο

$$H_0 : M = 5 \quad - \quad H_1 : M \neq 5 .$$

Αρχικά σχηματίζουμε τις διαφορές $D_i = X_i - 5$ που είναι οι ακόλουθες

0, 4, 5, 3, -1, 3, 0, -2, -5, 5, 10, 4, 0.

Παρατηρούμε ότι τρεις από αυτές είναι 0, οπότε τις αποβάλλουμε, και κατασκευάζουμε τον ακόλουθο πίνακα

i	X_i	D_i	$\lvert D_i \rvert$	R_i	k_i	$k_i R_i$	$(1-k_i)R_i$
1	9	4	4	5.5	1	5.5	0
2	10	5	5	8	1	8	0
3	8	3	3	3.5	1	3.5	0
4	4	-1	1	1	0	0	1
5	8	3	3	3.5	1	3.5	0
6	3	-2	2	2	0	0	2
7	0	-5	5	8	0	0	8
8	10	5	5	8	1	8	0
9	15	10	10	10	1	10	0
10	9	4	4	5.5	1	5.5	0
				55		$t^+ = 44$	$t^- = 11$

Συνεπώς, για τον προσεγγιστικό έλεγχο έχουμε $n = 10$, $t^+ = 44$ και

$$t^* = \frac{t^+ - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{3}(t_j^3 - t_j)}} = \frac{44 - 27.5}{\sqrt{96.25 - \frac{(2^3-2)+(2^3-2)+(3^3-3)}{48}}} = 1.68842688 .$$

Συνεπώς $p-value = 2(1-\Phi(|t^*|)) = 0.09132931$.

Χρησιμοποιώντας το R παίρνουμε

```
> x <- c(5,9,10,8,4,8,5,3,0,10,15,9,5)
> wilcox.test(x, mu=5, alternative="t", correct=FALSE, exact=FALSE)

        Wilcoxon signed rank test

data:  x
V = 44, p-value = 0.09133
alternative hypothesis: true location is not equal to 5
```

Αξίζει να σημειώνουμε ότι για τον παραπάνω έλεγχο θα μπορούσε να χρησιμοποιηθεί και το `t-test` αφού η υπόθεση της κανονικότητας των παρατηρήσεων δεν απορρίπτεται (αυτό το σημείο θα εξεταστεί αργότερα).

```
> t.test(x, mu=5)

        One Sample t-test

data:  x
t = 1.8723, df = 12, p-value = 0.08572
alternative hypothesis: true mean is not equal to 5
```

Κλείνοντας σημειώνουμε ότι ο έλεγχος Wilcoxon Signed-Rank για ένα δείγμα μπορεί να χρησιμοποιηθεί και ως έλεγχος για τη συμμετρικότητα ενός πληθυσμού, όπως ορισμένοι συγγραφείς προτείνουν. Η μηδενική υπόθεση είναι ότι (α) ο πληθυσμός είναι συμμετρικός, και (β) η διάμεσος (ή η μέση τιμή) είναι ίση με θ_0.

Παράδειγμα 6.9 (συμμετρικότητα πληθυσμού)

Για παράδειγμα ας θεωρήσουμε την εκθετική κατανομή (που δεν είναι συμμετρική) με παράμετρο $\lambda = 5$ και διάμεσο $M_0 = (\log 2)/5$. Ο έλεγχος Wilcoxon Signed-Rank και ο προσημικός έλεγχος δίνουν τελείως διαφορετικά συμπεράσματα.

```
> x <-rexp(500,rate=5); Med <- log(2)/5
> wilcox.test(x, mu=Med)

        Wilcoxon signed rank test with continuity correction

data:  x
V = 78370, p-value = 1.111e-06
alternative hypothesis: true location is not equal to 0.1386294

> SIGN.test(x, md=Med)

        One-sample Sign-Test

data:  x
s = 253, p-value = 0.8231
alternative hypothesis: true median is not equal to 0.1386294
```

Ο έλεγχος Wilcoxon Signed-Rank λανθασμένα απορρίπτει τη μηδενική υπόθεση (αυτό δικαιολογείται από το γεγονός ότι δεν ικανοποιείται η υπόθεση της συμμετρικότητας του πληθυσμού), ενώ ο προσημικός έλεγχος σωστά δεν απορρίπτει τη μηδενική υπόθεση. Στη γενική περίπτωση η απόρριψη της μηδενικής υπόθεσης σημαίνει ότι ο πληθυσμός δεν είναι συμμετρικός είτε/και η διάμεσος δεν είναι ίση με την προκαθορισμένη τιμή.

Τα συμπεράσματα όμως για την ομοιόμορφη κατανομή $U(0,5)$ (συμμετρική κατανομή) με $M_0 = 2.5$ θα πρέπει να είναι λογικά τα ίδια. Πράγματι

```
> x <-runif(500, min=0, max=5)
> wilcox.test(x, mu=2.5)

        Wilcoxon signed rank test with continuity correction

data:  x
V = 66883, p-value = 0.1878
alternative hypothesis: true location is not equal to 2.5

> SIGN.test(x, md=2.5)

        One-sample Sign-Test
```

```
data:  x
s = 266, p-value = 0.1656
alternative hypothesis: true median is not equal to 2.5
```

Επομένως σε αυτή την περίπτωση το συμπέρασμα του Wilcoxon Signed-Rank ελέγχου είναι ότι ο πληθυσμός είναι συμμετρικός με διάμεσο 2.5. ■

6.2.6 Έλεγχος τυχαιότητας με το κριτήριο των ροών (Wald-Wolfowitz)

Όλοι οι έλεγχοι που συναντήσαμε μέχρι τώρα βασίζονταν στην υπόθεση ότι το δείγμα μας είναι τυχαίο. Στην παρούσα παράγραφο θα παρουσιάσουμε ένα έλεγχο τυχαιότητας του δείγματος χρησιμοποιώντας το κριτήριο του αριθμού των ροών που οφείλεται στους Wald και Wolfowitz. Αν παρατηρούσαμε την ακόλουθη σειρά των ανδρών και των γυναικών (Α: άνδρας, Γ: γυναίκα)

Α Γ Α Γ Α Γ Α Γ Α Γ

που περιμένουν στο ταμείο ενός σινεμά για να αγοράσουν εισιτήριο δεν θα μπορούσαμε να υποστηρίξουμε ότι η συγκεκριμένη σειρά είναι τυχαία. Στο ίδιο συμπέρασμα θα καταλήγαμε παρατηρώντας και τη σειρά

Α Α Α Α Α Γ Γ Γ Γ Γ.

Στην πρώτη σειρά υπάρχουν συνολικά 10 ροές (πλήθος από ομάδες συνεχόμενων όμοιων συμβόλων) ενώ στη δεύτερη μόνο δύο. Επομένως φαίνεται λογικό να αναπτυχθεί ένας έλεγχος τυχαιότητας που ως στατιστική συνάρτηση ελέγχου θα χρησιμοποιεί το συνολικό αριθμό R των ροών μιας ακολουθίας αποτελεσμάτων δύο συμβόλων. Ένας τέτοιος έλεγχος είναι λογικό να απορρίπτει τη μηδενική υπόθεση της τυχαιότητας της ακολουθίας έναντι της ("δίπλευρης") εναλλακτικής υπόθεσης της μη-τυχαιότητας, όταν υπάρχει υπερβολικά μικρός ή υπερβολικά μεγάλος αριθμός ροών. Η εναλλακτική υπόθεση όμως θα μπορούσε να είναι και μονόπλευρη. Αν η εναλλακτική υπόθεση υποστηρίζει ότι υπάρχει κάποιου είδους τάση να "ανακατώνονται" τα δύο σύμβολα τότε η μηδενική απόθεση θα απορρίπτεται προς χάριν της εναλλακτικής για μεγάλες τιμές του R (όπως συμβαίνει στην πρώτη περίπτωση). Αν η εναλλακτική υπόθεση υποστηρίζει ότι υπάρχει κάποιου είδους τάση "ομαδοποίησης" στα όμοια σύμβολα τότε η μηδενική απόθεση θα απορρίπτεται προς χάριν της εναλλακτικής για μικρές τιμές του R (όπως συμβαίνει στη δεύτερη περίπτωση).

Ο έλεγχος τυχαιότητας που περιγράφτηκε παραπάνω μπορεί να χρησιμοποιηθεί για οποιοδήποτε δείγμα $X_1, X_2, ..., X_n$ (ο υποδείκτης κάθε παρατήρησης δηλώνει τη σειρά με την οποία συλλέχτηκαν οι n παρατηρήσεις), και όχι μόνο στην περίπτωση που τα X_i είναι δίτιμες τυχαίες μεταβλητές. Απλά αντιστοιχούμε σε κάθε X_i τις τιμές 0 ή 1 (ή ακόμη και τις τιμές -1, 1) ανάλογα με το αν η παρατήρηση X_i είναι μικρότερη ή μεγαλύτερη από τη διάμεσο (ή τη μέση τιμή) του δείγματος και στη συνέχεια εφαρμόζουμε το κριτήριο των ροών στην ακολουθία των δύο συμβόλων που προέκυ-

ψε. Αν κάποια παρατήρηση είναι ίση με τη διάμεσο (ή τη μέση τιμή) τότε την αποβάλλουμε και μειώνουμε το μέγεθος του δείγματος ισόποσα.

Για τη μαθηματική διατύωση του κριτηρίου των ροών ας θεωρήσουμε μια ακολουθία n συμβόλων δύο τύπων, όπου n_1 είναι το πλήθος των συμβόλων του ενός τύπου και n_2 είναι το πλήθος των συμβόλων του άλλου τύπου ($n = n_1 + n_2$). Η κατανομή του αριθμού R των ροών σε μια τέτοιου είδους ακολουθία είναι φυσικά διακριτή με τιμές 2, 3, ..., $n_1 + n_2$. Για την ακριβή μορφή της κατανομής του R, όταν ισχύει η μηδενική υπόθεση, παραπέμπουμε σε βιβλία μη παραμετρικής στατιστικής. Ωστόσο για μεγάλα n_1 και n_2 (αμφότερα μεγαλύτερα του 12) μπορούμε να υπολογίσουμε προσεγγιστικά πιθανότητες που αφορούν το R χρησιμοποιώντας την ασυμπτωτική κατανομή της τυχαίας μεταβλητής

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}} = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$$

που είναι η $N(0,1)$ όταν η μηδενική υπόθεση είναι αληθής.

Αν συμβολίσουμε με *r* και με *z* την παρατηρούμενη τιμή της *R* και της *Z*, αντιτοίχως, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η (προσεγγιστική) $p-value$ του ελέγχου δίνονται στον ακόλουθο πίνακα

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$p-value$
Ομαδοποίηση `positive.correlated`	$\{r : r \le r_a\}$	$P(Z \le z)$
Ανακάτωμα `negative.correlated`	$\{r : r \ge r'_a\}$	$P(Z \ge z)$
Μη-τυχαιότητα `two.sided`	$\{r : r \le r_{a/2}\} \bigcup \{r : r \ge r'_{a/2}\}$	$2\min\{P(Z \le z), P(Z \ge z)\}$

Οι τιμές r_a και r'_a που εμφανίζονται στο παραπάνω πλαίσιο υπολογίζονται κάνοντας χρήση της ακριβής κατανομής του *R* σύμφωνα με τις σχέσεις $P(R \le r_a) \le a$ και $P(R \ge r'_a) \le a$.

Στο R ο έλεγχος τυχαιότητας με το κριτήριο των ροών πραγματοποιείται με τη συνάρτηση `runs.test` του πακέτου `lawstat` (δείτε επίσης την ομώνυμη συνάρτηση του πακέτου `tseries`). Η βασική σύνταξη της συνάρτησης `runs.test` είναι η ακόλουθη

```
runs.test(x, alternative=c("two.sided", positive.correlated",
    "negative.correlated")
x: το δείγμα σε μορφή διανύσματος
```

Παράδειγμα 6.10 (προσεγγιστικός έλεγχος – συνεχής πληθυσμός)

Θα χρησιμοποιήσουμε τα δεδομένα του Παραδείγματος 6.3 για να ελέγξουμε αν το δείγμα με τα 30 βάρη X_i $(i = 1, 2, ..., 30)$ των σάκων με τσιμέντο είναι τυχαίο σε επίπεδο σημαντικότητας $a = 0.1$. Αρχικά υπολογίζουμε τη (δειγματική) διάμεσο των δεδομένων που είναι ίση με 25.25, στη συνέχεια υπολογίζουμε τις διαφορές $X_i - 25.25$, και δημιουργούμε μια νέα μεταβλητή την Y_i ως εξής

$$Y_i = \begin{cases} -1, & X_i - 25.25 < 0 \\ \\ 1, & X_i - 25.25 > 0. \end{cases}$$

Οι υπολογισμοί δίνονται στον ακόλουθο πίνακα

i	X_i	$X_i - 25.25$	Y_i	i	X_i	$X_i - 25.25$	Y_i
1	26.18	0.93	1	16	26.24	0.99	1
2	25.30	0.05	1	17	25.46	0.21	1
3	25.18	-0.07	-1	18	25.01	-0.24	-1
4	24.54	-0.71	-1	19	24.71	-0.54	-1
5	25.14	-0.11	-1	20	25.27	0.02	1
6	25.44	0.19	1	21	24.22	-1.03	-1
7	24.49	-0.76	-1	22	24.49	-0.76	-1
8	25.01	-0.24	-1	23	25.68	0.43	1
9	25.12	-0.13	-1	24	26.01	0.76	1
10	25.67	0.42	1	25	25.50	0.25	1
11	24.22	-1.03	-1	26	25.84	0.59	1
12	26.48	1.23	1	27	26.09	0.84	1
13	23.97	-1.28	-1	28	25.21	-0.04	-1
14	25.83	0.58	1	29	26.04	0.79	1
15	25.05	-0.20	-1	30	25.23	-0.02	-1

Στην ακολουθία Y_1, Y_2, ..., Y_{30} υπάρχουν $n_1 = 15$ "1", $n_2 = 15$ "-1" ενώ το $r = 18$. Επειδή $n_1, n_2 > 12$ θα χρησιμοποιήσουμε τον προσεγγιστικό έλεγχο. Η παρατηρούμενη τιμή της στατιστική συνάρτηση ελέγχου Z είναι ίση με

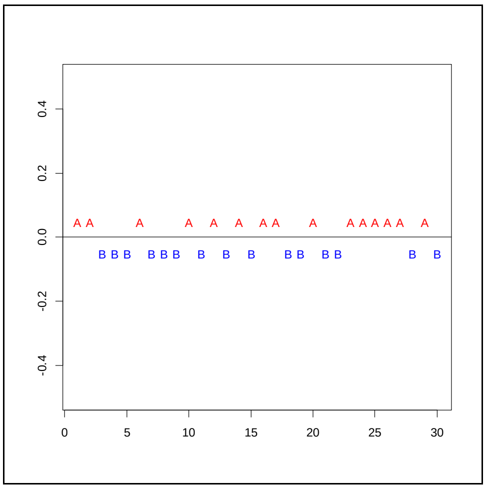
$$z = \frac{r - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}} = \frac{18 - \left(\frac{2 \cdot 15 \cdot 15}{15 + 15} + 1 \right)}{\sqrt{\frac{2 \cdot 15 \cdot 15 (2 \cdot 15 \cdot 15 - 15 - 15)}{(15 + 15)^2 (15 + 15 - 1)}}} = 0.743223353$$

με $p - value = 2 \min\{P(Z \le z), P(Z \ge z)\} = 2 \cdot 0.2286732 = 0.4573465$. Επομένως συμπεραίνουμε ότι το δείγμα είναι τυχαίο (η μηδενική υπόθεση δεν απορρίπτεται).

Χρησιμοποιώντας το R παίρνουμε

```
> library(lawstat)
> cement <- read.table("cement.txt", header=T)
> attach(cement); names(cement)
 [1] "x"
> runs.test(x, plot.it=TRUE)
```

```
        Runs Test - Two sided

data:  x
Standardized Runs Statistic = 0.7432, p-value = 0.4573
```



Αξίζει να σημειώνουμε ότι η συνάρτηση `runs.test` του πακέτου `lawstat` δουλεύει και για δίτιμα δεδομένα. Για την επαλήθευση του παραπάνω ελέγχου με δίτιμα δεδομένα έχουμε:

```
> xnew <- ifelse (x<median(x), -1, 1)
> xnew
 [1]  1  1 -1 -1 -1  1 -1 -1 -1  1 -1  1 -1  1 -1
[16]  1  1 -1 -1  1 -1 -1  1  1  1  1  1 -1  1 -1
> runs.test(xnew)

        Runs Test - Two sided

data:  xnew
Standardized Runs Statistic = 0.7432, p-value = 0.4573
```

Κλείνοντας την παρούσα παράγραφο σημειώνουμε ότι ο συνολικός αριθμός των ροών σε μια ακολουθία συμβόλων δύο τύπων μπορεί να χρησιμοποιηθεί για να συμπερασματολογήσουμε για το αν δύο ανεξάρτητα τυχαία δείγματα $X_1, X_2, ..., X_{n_1}$ και $Y_1, Y_2, ..., Y_{n_2}$ προέρχονται από τον ίδιο πληθυσμό ή όχι. Μας ενδιαφέρει δηλαδή ο έλεγχος

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \neq F_Y(x)$$

(η H_0 ισχύει για όλα τα x ενώ η H_1 ισχύει για τουλάχιστον ένα x) όπου $F_X(x)$ και $F_Y(x)$ οι συναρτήσεις κατανομής των δύο πληθυσμών X και Y. Αν και τέτοιου είδους έλεγχοι που αφορούν δύο δείγματα θα μας απασχολήσουν στην επόμενη ενότητα, ο συγκεκριμένος έλεγχος κρίνεται σκόπιμο να παρουσιαστεί εδώ.

Για να εκτελέσουμε τον έλεγχο αρχικά διατάσσουμε τις $n_1 + n_2$ παρατηρήσεις και αντικαθιστούμε τις παρατηρήσεις που αντιστοιχούν στο πρώτο δείγμα με a και τις παρατηρήσεις που αντιστοιχούν

στο δεύτερο δείγμα με b. Στη συνέχεια καταγράφουμε το συνολικό αριθμό των ροών R στην ακολουθία που προέκυψε με τα δύο σύμβολα a και b. Είναι προφανές ότι μικρές τιμές του R παρέχουν ενδείξεις υπέρ της απόρριψης της μηδενικής υπόθεσης. Η περιοχή απόρριψης του ελέγχου σε επίπεδο σημαντικότητας a είναι η $\{r : r \le r_a\}$ όπου το r_a βρίσκεται από τη σχέση $P(R \le r_a) \le a$.

6.2.7 Έλεγχος κανονικότητας ενός πληθυσμού

Σε αρκετούς ελέγχους που συναντήσαμε έως τώρα υπήρχε μια βασική προϋπόθεση για την εφαρμογή τους: η κανονικότητα του πληθυσμού. Στην παρούσα παράγραφο δίνονται έλεγχοι για να αξιολογήσουμε την υπόθεση της κανονικότητας του πληθυσμού.

6.2.7.1 Έλεγχος Kolmogorov-Smirnov

Έστω $X_1, X_2, ..., X_n$ τυχαίο δείγμα από πληθυσμό X με συνάρτηση κατανομής $F(x)$. Για τον έλεγχο

$$H_0 : F(x) = F_0(x) \quad - \quad H_1 : F(x) \neq F_0(x)$$

(η H_0 ισχύει για όλα τα x ενώ η H_1 ισχύει για τουλάχιστον ένα x) όπου $F_0(x)$ είναι μια πλήρως καθορισμένη συνάρτηση κατανομής συνεχούς τυχαίας μεταβλητής χρησιμοποιείται συνήθως ο έλεγχος Kolmogorov-Smirnov. Η στατιστική συνάρτηση D_n του ελέγχου δίνεται από τη σχέση

$$D_n = \sup_{-\infty < x < \infty} \{| S_n(x) - F_0(x) |\}$$

όπου $S_n(x)$ [†] είναι η εμπειρική συνάρτηση κατανομής του δείγματος. Προφανώς μεγάλες τιμές της D_n προσφέρουν ενδείξεις για την ισχύ της εναλλακτικής υπόθεσης. Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$$\{d_n : \ d_n > D_{n;\,a}\}$$

όπου d_n είναι η παρατηρούμενη τιμή της D_n, και $D_{n,a}$ το άνω a ποσοστιαίο σημείο της κατανομής του D_n. Η $p-value$ του ελέγχου είναι ίση με $P(D_n > d_n)$. Επειδή η ακριβής κατανομή της D_n είναι δύσκολο να υπολογιστεί, και για να διευκολυνθεί η εκτέλεση του ελέγχου, έχουν κατασκευαστεί ειδικοί πίνακες με τα ποσοστιαία σημεία της D_n. Ωστόσο, μπορεί να αποδειχθεί ότι

$$\lim_{n\to\infty} P(D_n \le d/\sqrt{n}) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

[†] $S_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$

οπότε μια προσεγγιστική τιμή της $p-value$ του ελέγχου για μεγάλα δείγματα (συνήθως $n>100$) είναι η

$$p-value = P(D_n > d_n) = 1 - P(D_n \le d_n) = 1 - P\left(D_n \le \frac{\sqrt{n} d_n}{\sqrt{n}} \right) \cong 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2ni^2 d_n^2} .$$

Ο έλεγχος Kolmogorov-Smirnov χρησιμοποιείται (όχι συχνά) και με μονόπλευρες εναλλακτικές υποθέσεις. Έτσι για τους ελέγχους

$$H_0 : F(x) = F_0(x) \quad - \quad H_{1,+} : F(x) \ge F_0(x),$$

$$H_0 : F(x) = F_0(x) \quad - \quad H_{1,-} : F(x) \le F_0(x),$$

(η H_0 ισχύει για όλα τα x ενώ στις $H_{1,+}$ και $H_{1,-}$ ισχύει γνήσια ανισότητα για ένα τουλάχιστον x) χρησιμοποιούνται, αντίστοιχα, οι ακόλουθες στατιστικές συναρτήσεις ελέγχου

$$D_n^+ = \sup_{-\infty<x<\infty} [S_n(x) - F_0(x)], \qquad D_n^- = \sup_{-\infty<x<\infty} [F_0(x) - S_n(x)]$$

(σημειώνεται ότι $D_n = \max\{D_n^+, D_n^-\}$). Οι κρίσιμες περιοχές δίνονται, αντίστοιχα, από τις σχέσεις

$$\{d_n^+ :\ d_n^+ > D_{n;\,a}^+\}, \qquad \{d_n^- :\ d_n^- > D_{n;\,a}^-\}$$

όπου d_n^+ και d_n^- είναι οι παρατηρούμενες τιμές των D_n^+ και D_n^-. Τα $D_{n,a}^+$ και $D_{n,a}^-$ δηλώνουν τα άνω a ποσοστιαία σημεία των κατανομών των D_n^+ και D_n^-, αντίστοιχα. Οι $p-value$ των ελέγχων δίνονται, αντίστοιχα, από τις σχέσεις $P(D_n^+ > d_n^+)$ και $P(D_n^- > d_n^-)$.

Αν και ο έλεγχος Kolmogorov-Smirnov έχει σχεδιαστεί για συνεχείς πληθυσμούς μπορεί να χρησιμοποιηθεί και για διακριτούς πληθυσμούς στους οποίους φυσικά η $F_0(x)$ είναι συνάρτηση κατανομής διακριτής τυχαίας μεταβλητής. Ωστόσο σε τέτοιες περιπτώσεις μπορούμε κάλλιστα να χρησιμοποιήσουμε τον έλεγχο καλής προσαρμογής χ^2 που θα αναπτυχθεί σε επόμενη παράγραφο.

Για την εκτέλεση του ελέγχου Kolmogorov-Smirnov στο R χρησιμοποιείται η συνάρτηση `ks.test`. Η βασική σύνταξη της συνάρτησης `ks.test` για ένα δείγμα είναι η ακόλουθη

```
ks.test(x, "string", ... , exact = TRUE or FALSE)
x: το δείγμα σε μορφή διανύσματος
"string": η συνάρτηση κατανομή της μηδενικής υπόθεσης (π.χ. pnorm,
    pt, punif)
... : οι παράμετροι της κατανομής της μηδενικής υπόθεσης
exact: p-value ακριβής (TRUE) ή προσεγγιστική (FALSE)
```

Παράδειγμα 6.11 (ks.test - δίπλευρη εναλλακτική)

Για να απαντήσουμε αν το τυχαίο δείγμα 5, 6, 7, 8, 9 προέρχεται από την κατανομή $N(7,1)$ εκτελούμε την ακόλουθη ανάλυση.

```
> x <- c(5,6,7,8,9)
-------------------------------------------------------------------------
> ks.test(x, "pnorm", mean=7, sd=1, exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2413, p-value = 0.9328
alternative hypothesis: two-sided
-------------------------------------------------------------------------
> ks.test(x, "pnorm", mean=7, sd=1, exact = TRUE)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2413, p-value = 0.8704
alternative hypothesis: two-sided
```

Για να διερευνήσουμε περαιτέρω τον έλεγχο Kolmogorov-Smirnov δίνουμε το ακόλουθο παράδειγμα που αναφέρεται στην ομοιόμορφη κατανομή $U(0,5)$.

```
> x <- runif(150, min=0, max=5)
-------------------------------------------------------------------------
> ks.test(x, "punif", min=0.5, max=4.5, exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.172, p-value = 0.00028
alternative hypothesis: two-sided
-------------------------------------------------------------------------
> ks.test(x, "punif", min=0, max=5, exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0963, p-value = 0.1241
alternative hypothesis: two-sided
```

Σε επίπεδο σημαντικότητας $a = 0.1$, ο πρώτος έλεγχος σωστά απορρίπτει τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από την κατανομή $U(0.5, 4.5)$, ενώ ο δεύτερος σωστά δεν απορρίπτει τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από την κατανομή $U(0,5)$. ■

Ο πρώτος έλεγχος του παραπάνω παραδείγματος αναφερόταν στην κανονική κατανομή $N(7,1)$. Ωστόσο, στην πράξη η μέση τιμή του πληθυσμού ή/και η διακύμανσή του δεν είναι συνήθως γνωστή. Μια καλή τακτική τότε θα ήταν να χρησιμοποιήσουμε στη θέση τους τις δειγματικές τιμές. Όμως σε αυτή την περίπτωση έχει παρατηρηθεί ότι ο έλεγχος γίνεται συντηρητικός (δηλαδή η ισχύς του ελέγχου είναι μικρότερη από την ισχύ του ελέγχου με γνωστές παραμέτρους). Μια καλή εναλλακτική λύση σε αυτή την περίπτωση είναι να εκτελεστεί ο έλεγχος του Lilliefors για τον έ-

λεγχο της κανονικότητας των δεδομένων που αποτελεί μια παραλλαγή του ελέγχου Kolmogorov-Smirnov (για τον έλεγχο του Lilliefors δείτε την επόμενη παράγραφο).

Παράδειγμα 6.12 (ks.test - μονόπλευρη εναλλακτική)

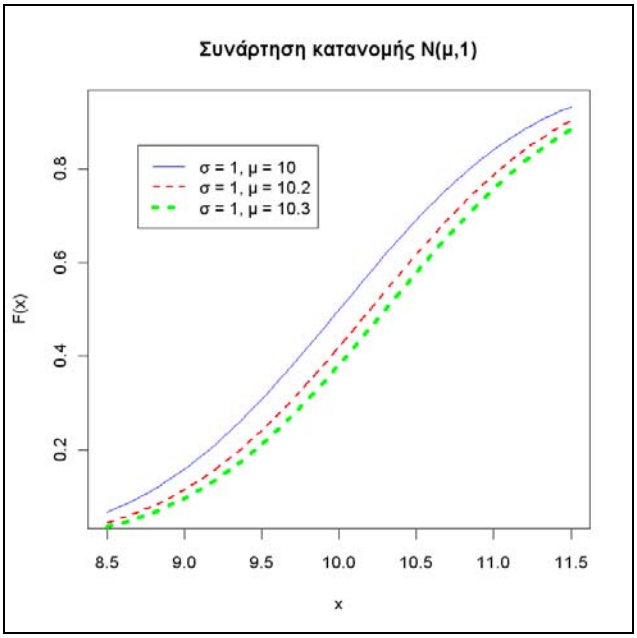
Για να διερευνήσουμε τον έλεγχο Kolmogorov-Smirnov με μονόπλευρες εναλλακτικές υποθέσεις δίνουμε το ακόλουθο παράδειγμα που αναφέρεται στον έλεγχο υπόθεσης

$$H_0 : F(x) = F_0(x) \quad - \quad H_{1,+} : F(x) \geq F_0(x).$$

Τα παραδείγματα έχουν κατασκευαστεί με τέτοιο τρόπο ώστε να είναι αληθείς στην πραγματικότητα οι εναλλακτικές υποθέσεις. Επομένως η $p-value$ του ελέγχου θα πρέπει να είναι αρκετά μικρή.

```
> x <- rnorm(150, mean=10, sd=1)
-------------------------------------------------------------------------
> ks.test(x, "pnorm", mean=10.2, sd=1, alternative="gr", exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D^+ = 0.1198, p-value = 0.01352
alternative hypothesis: the CDF of x lies above the null hypothesis
-------------------------------------------------------------------------
> ks.test(x, "pnorm", mean=10.3, sd=1, alternative="gr", exact = FALSE)

        One-sample Kolmogorov-Smirnov test

data:  x
D^+ = 0.154, p-value = 0.0008155
alternative hypothesis: the CDF of x lies above the null hypothesis
```

Για του λόγου το αληθές δίνουμε το ακόλουθο γράφημα



6.2.7.2 Άλλοι έλεγχοι κανονικότητας ενός πληθυσμού

Πέρα από τον έλεγχο Komogorov-Smirnov για τη διερεύνηση της κανονικότητας ενός πληθυσμού υπάρχουν και άλλοι διαθέσιμοι έλεγχοι. Ίσως ο σημαντικότερος από αυτούς (έχει γενικά μεγάλη ισχύ) είναι ο έλεγχος Shapiro-Wilk. Ο έλεγχος Shapiro-Wilk προτείνεται γενικά για μικρά δείγματα ($n < 50$) αλλά συμπεριφέρεται καλά και για δείγματα με $n < 2000$ (για $n > 2000$ προτείνεται ο έλεγχος Kolmogorov-Smirnov αν και τότε οι διαφορές στην ισχύ μεταξύ των δύο τεστ είναι αμελητέες). Ο έλεγχος εκτελείται με τη συνάρτηση `shapiro.test`. Η στατιστική συνάρτηση ελέγχου αποτιμά πόσο κοντά είναι τα εμπειρικά ποσοστιαία σημεία του δείγματος από τα αντίστοιχα θεωρητικά ποσοστιαία σημεία μιας κατάλληλης κανονικής κατανομής.

Ένα μειονέκτημα του ελέγχου Komogorov-Smirnov, είναι ότι η κατανομή της στατιστικής συνάρτησης ελέγχου εξαρτάται από τις παραμέτρους του πληθυσμού. Έτσι αν δεν είναι εκ των προτέρων γνωστές οι παράμετροι και αναγκαστούμε να τις εκτιμήσουμε και να τις θέσουμε ως πραγματικές τιμές των παραμέτρων των αντίστοιχων ορισμάτων της συνάρτησης `ks.test`, τότε το αποτέλεσμα του ελέγχου δεν θα είναι αξιόπιστο. Σε αυτή την περίπτωση είναι προτιμότερο να χρησιμοποιήσουμε μια παραλλαγή του ελέγχου Komogorov-Smirnov, τον έλεγχο Lilliefors (χρησιμοποιεί προσομοίωση Monte Carlo) που αποδίδει μια προσαρμοσμένη $p-value$. Ο συγκεκριμένος έλεγχος εκτελείται με τη συνάρτηση `lillie.test` του πακέτου `nortest`.

Το πακέτο nortest περιέχει άλλα 4 βασικά τεστ για τον έλεγχο της κανονικότητας ενός τυχαίου δείγματος. Οι συναρτήσεις είναι οι εξής:

- `sf.test`: Shapiro-Francia test
- `ad.test`: Anderson-Darling test
- `cvm.test`: Cramer-von Mises test
- `pearson.test`: Pearson chi-square test

Επίσης ένας άλλος σημαντικός έλεγχος κανονικότητας ενός τυχαίου δείγματος με μέγεθος $n \geq 20$ είναι ο έλεγχος των D'Agostino-Pearson που χρησιμοποιεί το συντελεστή ασυμμετρίας και κύρτωσης των δεδομένων (πακέτο `fBasics`, συνάρτηση `dagoTest`).

Παράδειγμα 6.13 (ks.test έναντι lillie.test)

Για να απαντήσουμε αν το τυχαίο δείγμα 7, 8, 9, 10, 11, 12, 13, 16, 20, 25 προέρχεται από κανονική κατανομή εκτελούμε την ακόλουθη ανάλυση.

```
> x <- c(7,8,9,10,11,12,13,16,20,25)
-------------------------------------------------------------------------
> ks.test(x, "pnorm", mean=mean(x), sd=sd(x), exact=FALSE)
        One-sample Kolmogorov-Smirnov test
```

```
data:  x
D = 0.207, p-value = 0.7849
alternative hypothesis: two-sided
------------------------------------------------------------------------
> shapiro.test(x)
        Shapiro-Wilk normality test

data:  x
W = 0.8942, p-value = 0.1889
------------------------------------------------------------------------
> library(nortest)
> lillie.test(x)
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  x
D = 0.207, p-value = 0.2602
------------------------------------------------------------------------
> ad.test(x)
        Anderson-Darling normality test

data:  x
A = 0.4578, p-value = 0.2059
------------------------------------------------------------------------
> sf.test(x)
        Shapiro-Francia normality test

data:  x
W = 0.8954, p-value = 0.1651
------------------------------------------------------------------------
> cvm.test(x)
        Cramer-von Mises normality test

data:  x
W = 0.0764, p-value = 0.206
------------------------------------------------------------------------
> pearson.test(x)
        Pearson chi-square normality test

data:  x
P = 4.4, p-value = 0.2214
```

Παρατηρούμε ότι κανένα τεστ δεν απορρίπτει την υπόθεση της κανονικότητας του πληθυσμού. Ωστόσο δεν διαφεύγει από την προσοχή μας η αυξημένη τιμή της $p-value$ του ελέγχου Kolmogorov-Smirnov έναντι των υπόλοιπων ελέγχων. Ούτως ή άλλως δεν συνίσταται ο έλεγχος Komogorov-Smirnov αφού η κατανομή το πληθυσμού δεν καθορίζεται με ακρίβεια από τη μηδενική υπόθεση. Επίσης η παρατηρούμενη τιμή της στατιστικής συνάρτησης ελέγχου (`D = 0.207`) είναι ίδια στους ελέγχους Kolmogorov-Smirnov και Lilliefors.

6.2.7.3 Q-Q διάγραμμα

Το επονομαζόμενο Q-Q διάγραμμα είναι ένα διάγραμμα που μας επιτρέπει να αντιληφθούμε αν ένα τυχαίο δείγμα $X_1,\ X_2,\ ....,\ X_n$ προέρχεται από πληθυσμό με συνάρτηση κατανομής $F(x)$. Σε ένα Q-Q διάγραμμα απεικονίζονται τα σημεία $(F^{-1}(p_i), x_{(i)})$, $i=1,2,...,n,$ όπου

$$p_i = (i-a)/(n+1-2a), \quad a = \begin{cases} 3/8, & n \le 10 \\ 1/2, & n > 10 \end{cases}.$$

(το σημείο $F^{-1}(p_i)$ αποτελεί το (κάτω) p_i – ποσοστιαίο σημείο της κατανομής F). Αν τα σημεία του διαγράμματος αγκαλιάζουν την ευθεία $y=x$ τότε συμπεραίνουμε ότι το τυχαίο δείγμα προέρχεται από πληθυσμό με συνάρτηση κατανομής $F(x)$. Αν τα σημεία αγκαλιάζουν μια άλλη ευθεία τότε το τυχαίο δείγμα προέρχεται από μια ευρύτερη οικογένεια κατανομών που τα μέλη της διαφέρουν μόνο ως προς θέση ή την κλίμακα και ένας «τυπικός» αντιπρόσωπός της είναι η $F(x)$.

Για τη δημιουργία του παραπάνω Q-Q διαγράμματος χρησιμοποιούμε τη συνάρτηση `qqplot` με τα παρακάτω ορίσματα

```
qqplot(x, y)
x: τα pi ποσοστιαία σημεία της F
y: το δείγμα σε μορφή διανύσματος (οι παρατηρήσεις xi)
```

Σημειώνουμε ότι τα σημεία p_i, $i=1,2,...,n$, μπορούν να παραχθούν άμεσα με τη συνάρτηση `ppoints(n)`.

Ειδικότερα, για να ελεγχθεί αν τα δεδομένα προέρχονται από την κανονική κατανομή απεικονίζονται στο διάγραμμα τα σημεία $(\Phi^{-1}(p_i), x_{(i)})$, $i=1,2,...,n$. Εναλλακτικά μπορεί να χρησιμοποιηθεί η συνάρτηση `qqnorm` με την ακόλουθη σύνταξη

```
qqnorm(x)
x: το δείγμα σε μορφή διανύσματος
```

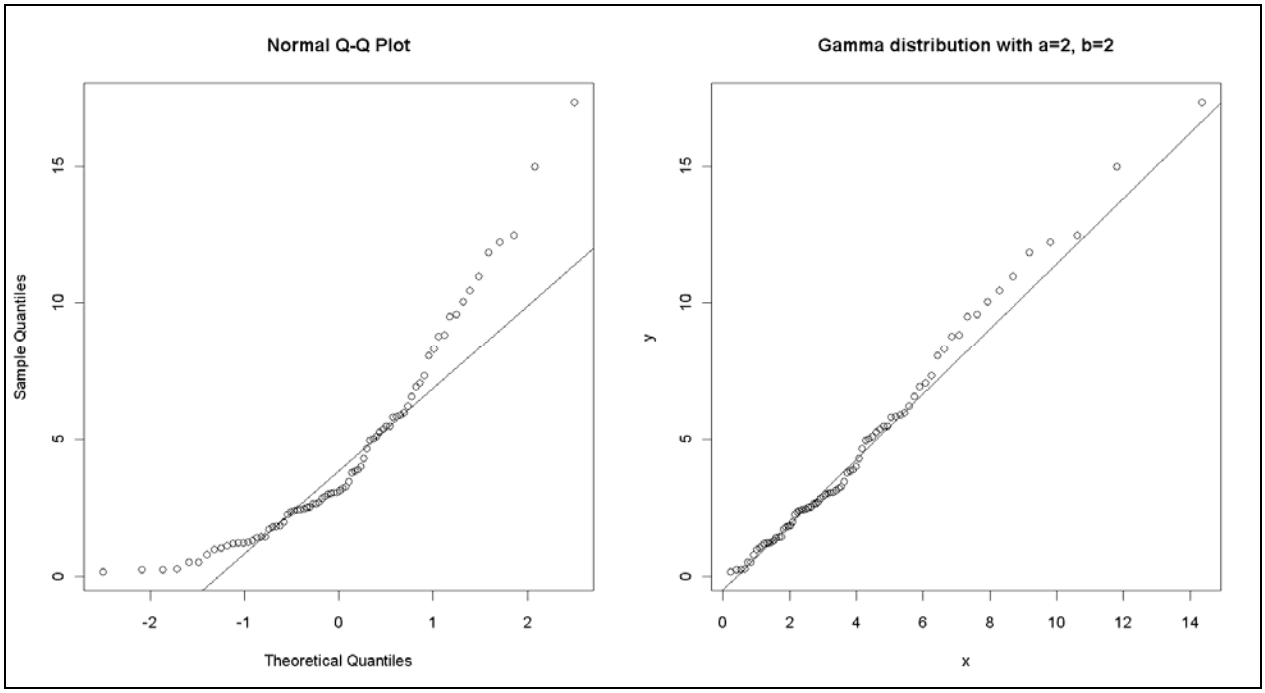
Όταν χρησιμοποιείται η συνάρτηση `qqnorm(x)`, η συνάρτηση `qqline(x)` προσθέτει μια ευθεία γραμμή στο κανονικό Q-Q διάγραμμα που περνά από τα σημεία

$(\Phi^{-1}(0.25), x_{0.25})$ και $(\Phi^{-1}(0.75), x_{0.75})$.

Παράδειγμα 6.14 (Q-Q διάγραμμα για την κατανομή G(2,2))

Ακολούθως παίρνουμε ένα δείγμα μεγέθους 80 από την κατανομή $G(2,2)$ και κατασκευάζουμε (α) το κανονικό Q-Q διάγραμμα, και (β) το Q-Q διάγραμμα που αντιστοιχεί στην κατανομή $G(2,2)$. Θεωρητικά τα σημεία του πρώτου διαγράμματος, σε αντίθεση με το δεύτερο, δεν πρέπει να βρίσκονται πάνω σε μια ευθεία γραμμή. Επίσης στη δεύτερη περίπτωση κατασκευάζεται η ευθεία που περνά από το πρώτο και το τρίτο τεταρτημόριο των δεδομένων.

```
> par(mfrow=c(1,2))
> y <- rgamma(80, shape=2, scale=2)
> qqnorm(y);qqline(y)
> x <-qgamma(ppoints(80),shape=2, scale=2)
> qqplot(x, y, main="Gamma distribution with a=2, b=2")
> a1 <- quantile(x, 0.25);b1 <- quantile(y, 0.25)
> a2 <- quantile(x, 0.75);b2 <- quantile(y, 0.75)
> abline(b1-a1*((b2-b1)/(a2-a1)),(b2-b1)/(a2-a1))
```

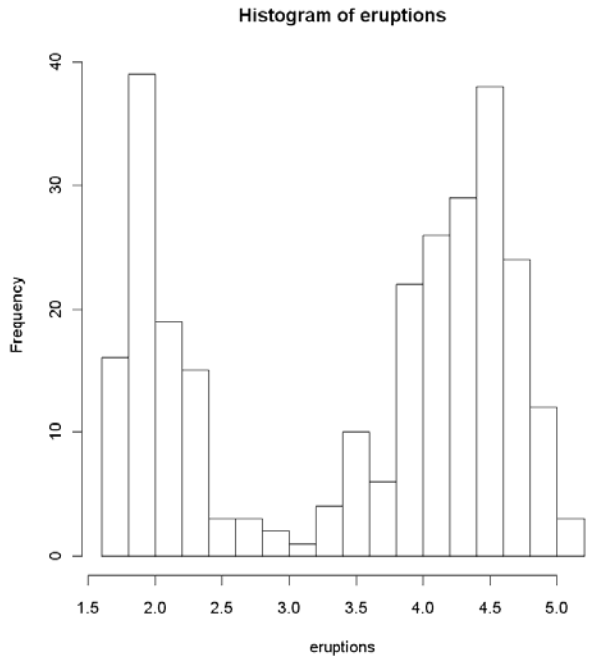


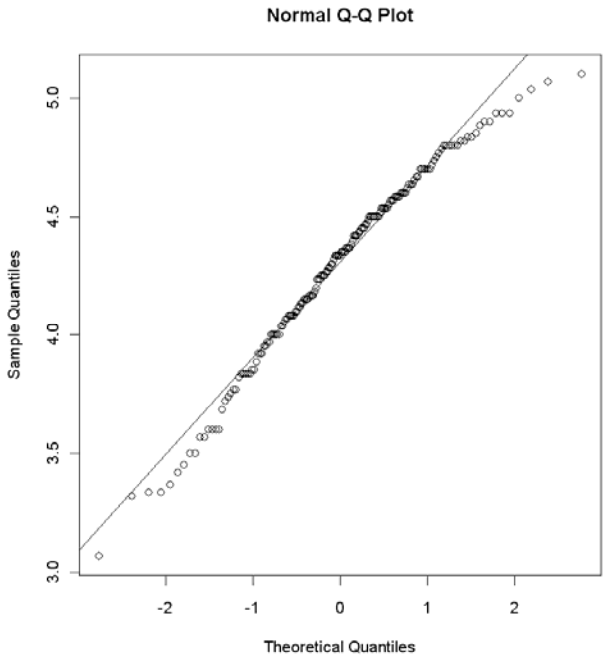
Στη συνέχεια δίνουμε άλλο ένα παράδειγμα που αφορά τα δεδομένα του αρχείου `faithful` του R για να διαφωτίσουμε τη λειτουργία της συνάρτησης `qqnorm`.

Παράδειγμα 6.15 (κανονικό Q-Q διάγραμμα)

Τα δεδομένα της μεταβλητής eruptions του αρχείου `faithful` φαίνονται να προέρχονται από τη "μίξη" δύο διαφορετικών κατανομών. Επικεντρωνόμαστε στην κατανομή που βρίσκεται προς τα δεξιά και κατασκευάζουμε το κανονικό Q-Q διάγραμμα.

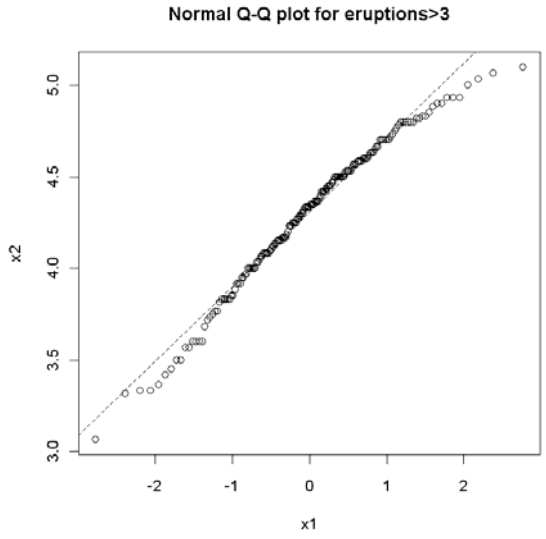
```
> attach(faithful)
> names(faithful)
[1] "eruptions" "waiting"
> par(mfrow=c(1,2))
> hist(eruptions, seq(1.6,5.2,0.2))
> long <- eruptions[eruptions>3]
> qqnorm(long);qqline(long)
```





Εναλλακτικά θα μπορούσαμε να κατασκευάσουμε το παραπάνω Q-Q διάγραμμα ως εξής

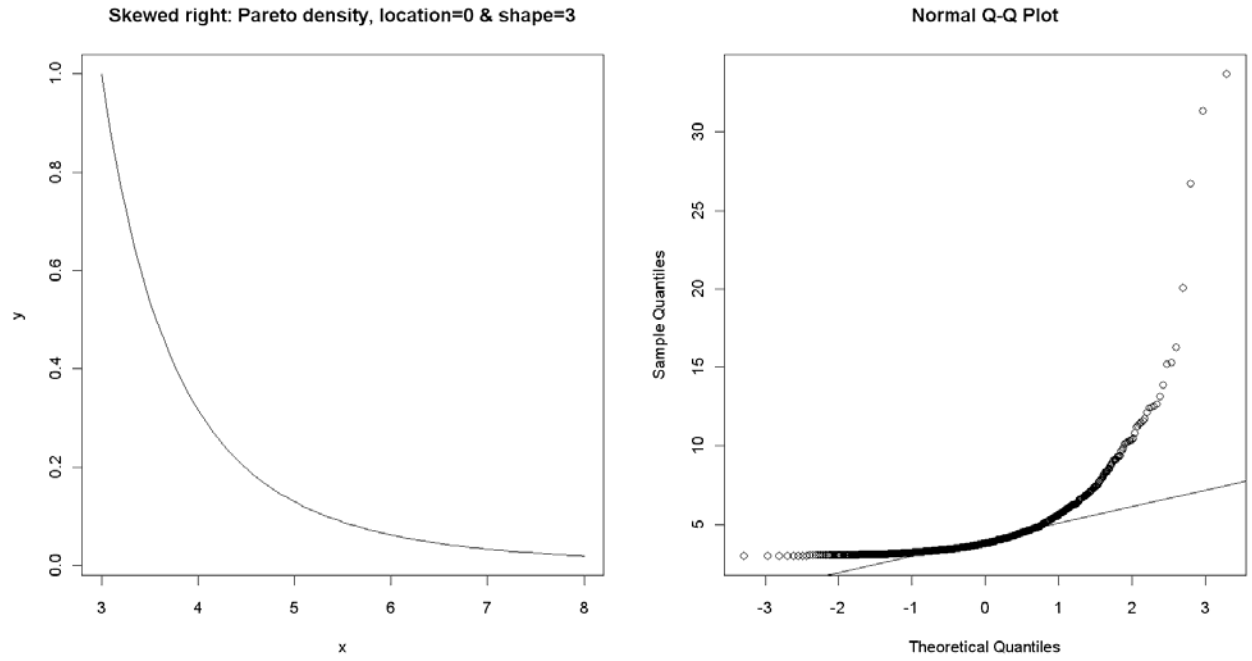
```
> n <- length(long)
> x1 <- qnorm((1:n-1/2)/n,0,1)
> x2 <- sort(long)
> plot(x1,x2); qqline(x2,lt=2)
```



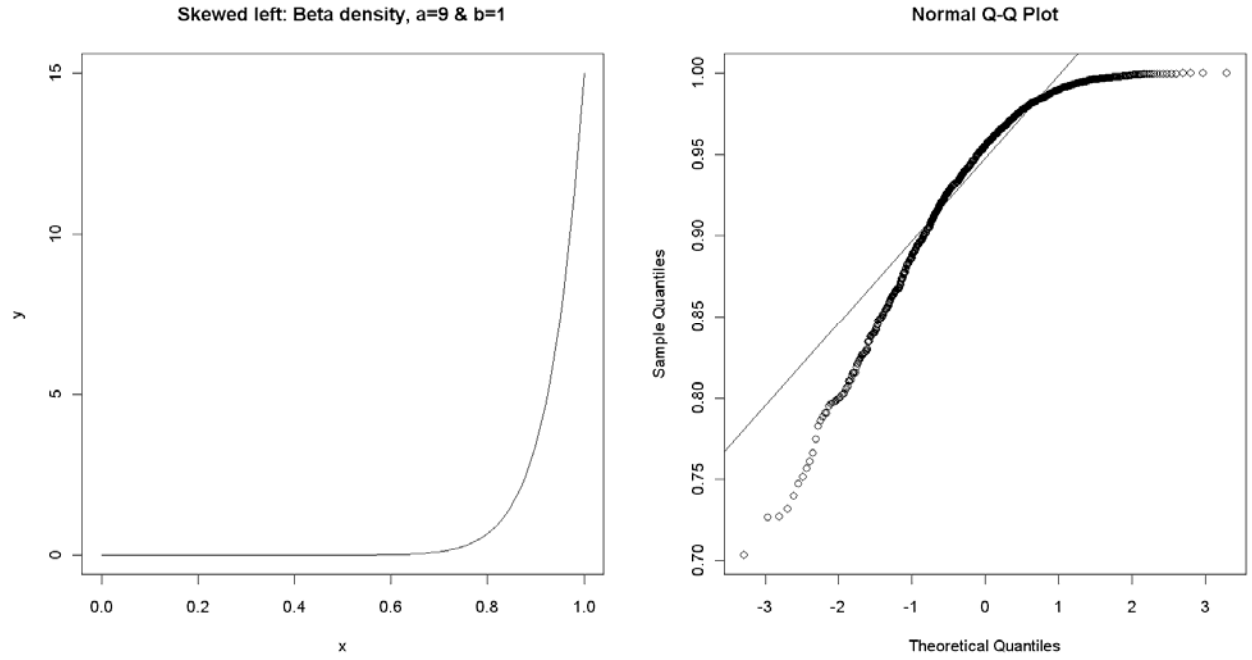
Παράδειγμα 6.16 (ερμηνεία κανονικών Q-Q διαγραμμάτων)

Παρακάτω δίνουμε βασικές μορφές κανονικών Q-Q διαγραμμάτων για κατανομές που είναι λοξές προς τα δεξιά (Pareto), λοξές προς τα αριστερά (Βήτα), πλατύκυρτες (ομοιόμορφη), λεπτόκυρτες (laplace), μεσόκυρτες (κανονική).

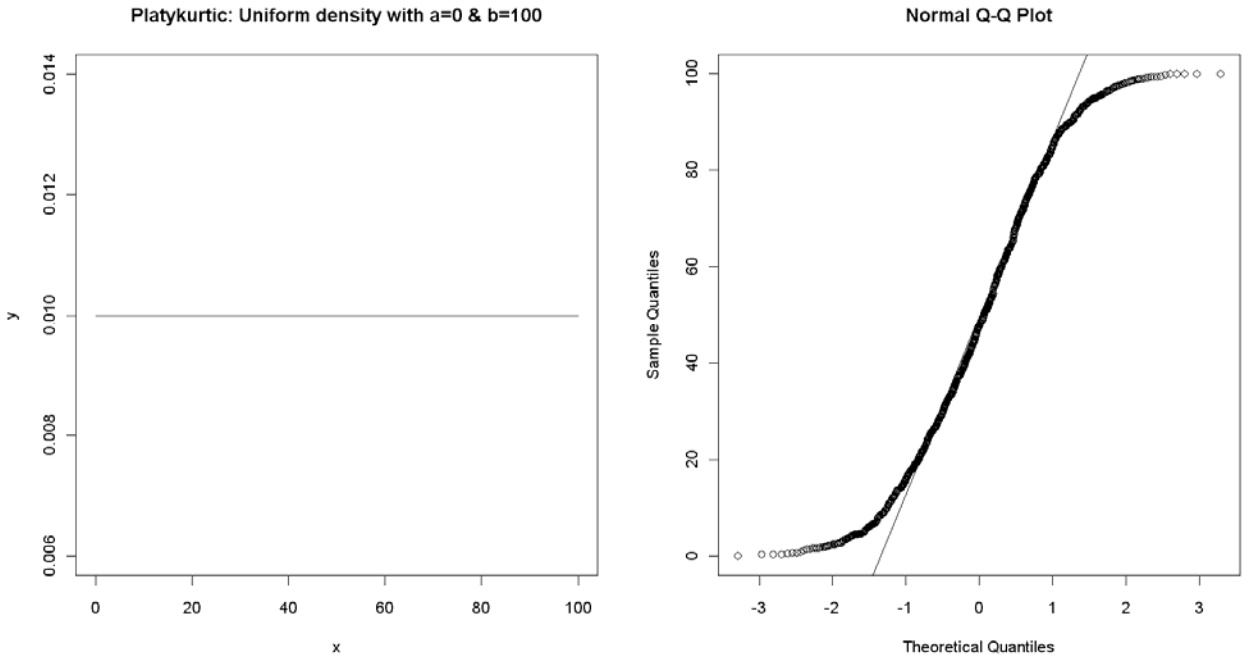
```
> library(VGAM)
> par(mfcol=c(1,2))
> alpha = 3; k = 3; x = seq(3.001, 8, length=100)
> y <- dpareto(x, location=alpha, shape=k)
> plot(x, y, type="l",main="Skewed right: Pareto density, location=0 &
shape=3")
> d <- rpareto(1000, location=alpha, shape=k);qqnorm(d);qqline(d)
```



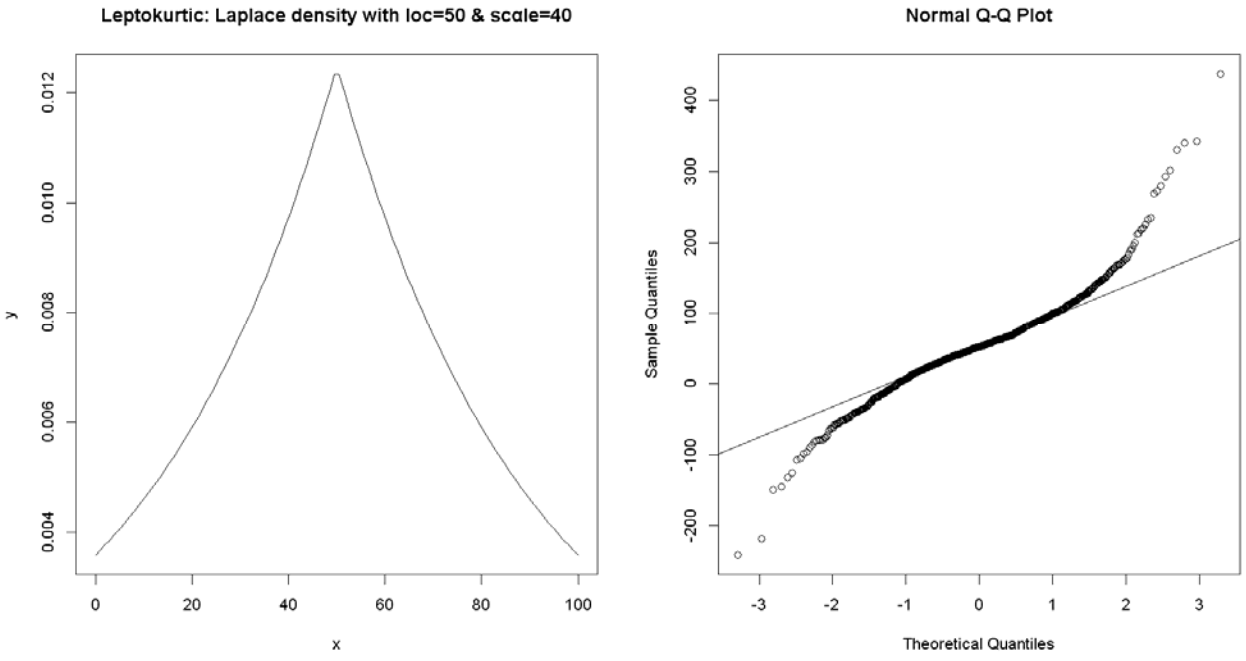

```
> par(mfcol=c(1,2))
> x <- seq(0, 1, length=100)
> y <- dbeta(x, shape1=15, shape2=1)
> plot(x, y,type="l", main="Skewed left: Beta density, a=9 & b=1")
> d <- rbeta(1000, shape1=15, shape2=1); qqnorm(d);qqline(d)
```



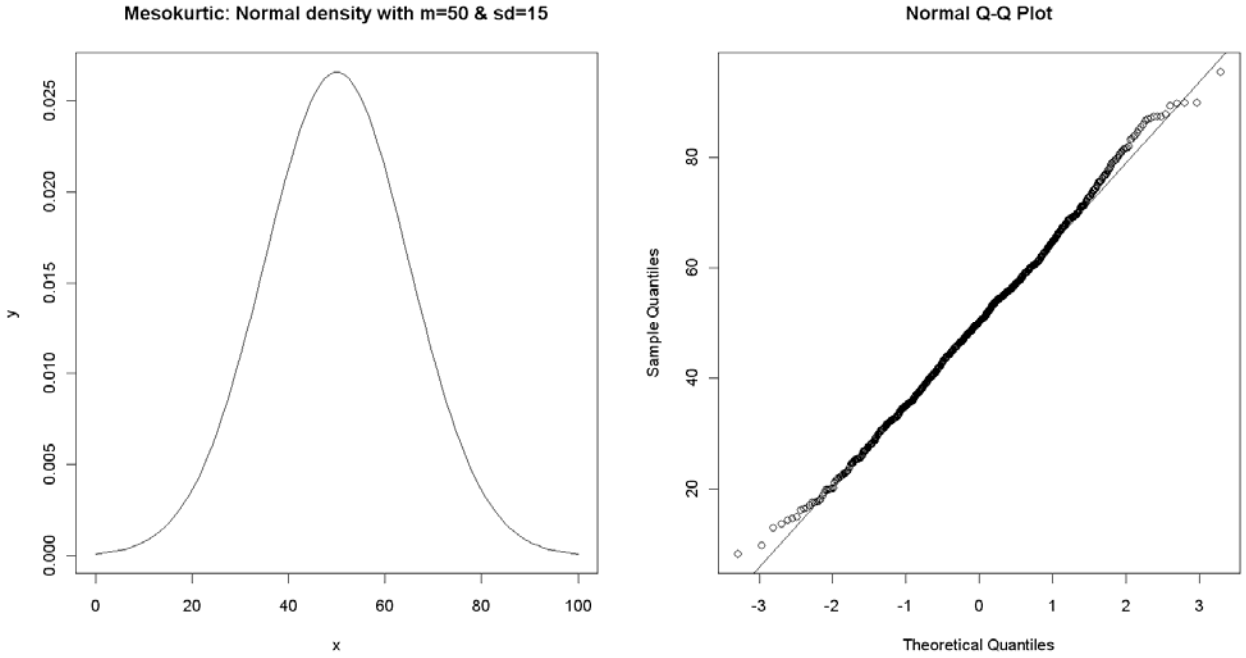
```
> par(mfcol=c(1,2))
> x <- seq(0, 100, length=100)
> y <- dunif(x, min=0, max=100)
> plot(x, y,type="l", main="Platykurtic: Uniform density with a=0 & b=100")
> d <- runif(1000, min=0, max=100); qqnorm(d); qqline(d)
```




```
> library(VGAM)
> par(mfcol=c(1,2))
> x <- seq(0, 100, length=100)
> y <- dlaplace(x, loc = 50, scale = 40)
> plot(x, y,type="l", main="Leptokurtic: Laplace density with loc=50 & scale=40")
> d <- rlaplace(1000, loc = 50, scale = 40); qqnorm(d); qqline(d)
```



```
> par(mfcol=c(1,2))
> x <- seq(0, 100, length=100)
> y <- dnorm(x, mean=50, sd = 15)
> plot(x, y,type="l", main="Mesokurtic: Normal density with m=50 &
sd=15")
> d <- rnorm(1000, mean=50, sd = 15); qqnorm(d); qqline(d)
```



6.3 Συμπερασματολογία για δύο δείγματα

6.3.1 Έλεγχος για τη διαφορά των μέσων τιμών δύο πληθυσμών (t-τεστ)

Έστω τα ανεξάρτητα δείγματα $X_1, X_2, ..., X_{n_1}$ και $Y_1, Y_2, ..., Y_{n_2}$ από κανονικούς πληθυσμούς $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα, με άγνωστες διακυμάνσεις. Για τον έλεγχο των υποθέσεων

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_1 - \mu_2 < \delta_0 ,$$

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_1 - \mu_2 > \delta_0 ,$$

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_1 - \mu_2 \neq \delta_0 ,$$

έχουμε τις ακόλουθες περιπτώσεις:

1η Περίπτωση: Για $\sigma_1^2 = \sigma_2^2$ (άγνωστα) η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\right)}} = \frac{\overline{X} - \overline{Y} - \delta_0}{S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

η οποία ακολουθεί την κατανομή $t_{n_1+n_2-2}$ όταν η H_0 είναι αληθής. Αν συμβολίσουμε με t την παρατηρούμενη τιμή της T, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η

$p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$\boldsymbol{p-value}$
$H_1 : \mu_1 - \mu_2 < \delta_0$	$\{t : t < -t_{n_1+n_2-2;\,a}\}$	$P(T \le t)$
$H_1 : \mu_1 - \mu_2 > \delta_0$	$\{t : t > t_{n_1+n_2-2;\,a}\}$	$P(T \ge t)$
$H_1 : \mu_1 - \mu_2 \ne \delta_0$	$\{t : t < -t_{n_1+n_2-2;\,a/2}\} \bigcup \{t : t > t_{n_1+n_2-2;\,a/2}\}$	$2P(T \ge \lvert t \rvert) = 2(1 - P(T \le \lvert t \rvert))$

Προτού εκτελεστεί ο παραπάνω έλεγχος πρέπει να προηγηθεί έλεγχος της ισότητας των διακυμάνσεων των δύο πληθυσμών (ο οποίος φυσικά δεν θα πρέπει να απορρίπτει την ισότητά τους). Ωστόσο ορισμένοι συγγραφείς θεωρούν ότι ο έλεγχος για την ισότητας των διακυμάνσεων των δύο πληθυσμών είναι περιττός όταν τα μεγέθη των δύο δειγμάτων να είναι ίσα.

Σημειώνεται επίσης ότι ο παραπάνω έλεγχος συνήθως χρησιμοποιείται στην πράξη όταν τα δείγματα είναι μικρού μεγέθους και οι κατανομές των δύο πληθυσμών δεν διαφέρουν πολύ από την κανονική κατανομή, αφού όταν τα μεγέθη και των δειγμάτων είναι μεγάλα δεν χρειάζεται να κάνουμε καμία περιοριστική υπόθεση ως προς την κατανομή των δύο δειγμάτων. Έτσι για μεγάλα δείγματα, όταν οι διασπορές είναι άγνωστες (αντίστοιχα γνωστές) χρησιμοποιείται η στατιστική συνάρτηση ελέγχου

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad \left(\text{αντίστοιχα} \quad Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a (και στις δύο περιπτώσεις) δίνεται στον ακόλουθο πίνακα ($w = t$ ή z)

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**
$H_1 : \mu_1 - \mu_2 < \delta_0$	$\{w : w < -z_a\}$
$H_1 : \mu_1 - \mu_2 > \delta_0$	$\{w : w > z_a\}$
$H_1 : \mu_1 - \mu_2 \ne \delta_0$	$\{w : w < -z_{a/2}\} \bigcup \{w : w > z_{a/2}\}$

2η Περίπτωση: Για $\sigma_1^2 \ne \sigma_2^2$ (άγνωστα) η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

η οποία ακολουθεί προσεγγιστικά την κατανομή t_ν όταν η H_0 είναι αληθής, όπου

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$p-value$
$H_1 : \mu_1 - \mu_2 < \delta_0$	$\{t : t < -t_{\nu;a}\}$	$P(T \le t)$
$H_1 : \mu_1 - \mu_2 > \delta_0$	$\{t : t > t_{\nu;a}\}$	$P(T \ge t)$
$H_1 : \mu_1 - \mu_2 \ne \delta_0$	$\{t : t < -t_{\nu;a/2}\} \bigcup \{t : t > t_{\nu;a/2}\}$	$2P(T \ge \lvert t \rvert) = 2(1 - P(T \le \lvert t \rvert))$

Σημειώνεται ότι ο έλεγχος της $2^{ης}$ Περίπτωσης χρησιμοποιείται στην πράξη όταν τα μεγέθη των δειγμάτων είναι μικρά και οι διακυμάνσεις δεν μπορούν να υποτεθούν ίσες.

Για τον έλεγχο της διαφοράς δύο μέσων τιμών μ_1 και μ_2 δύο πληθυσμών με ανεξάρτητα δείγματα χρησιμοποιείται η συνάρτηση `t.test` (δείτε επίσης Παράγραφο 6.2.1). Η βασική σύνταξη της συνάρτησης `t.test` για τον έλεγχο της διαφοράς των μέσων τιμών δύο πληθυσμών είναι η ακόλουθη

```
t.test(x, y, mu=δ0, var.equal=FALSE ή TRUE)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
mu: η τιμή της διαφοράς μ1-μ2 υπό τη μηδενική υπόθεση
var.equal: δηλώνει αν οι διακυμάνσεις είναι ίσες ή όχι
```

Παράδειγμα 6.17 (t.test)

Στο Παράδειγμα 6.1 δόθηκαν τα ύψη (σε cm) 60 τυχαία επιλεγμένων ανδρών από την περιοχή Α μιας χώρας. Στο ακόλουθο πλαίσιο δίνονται τα ύψη (σε cm) 40 τυχαία επιλεγμένων ανδρών από την περιοχή Β της χώρας (αρχείο HEIGHTB.txt)

176.6	177.5	166.8	172.0	175.0	177.7	178.5	177.7	178.6	178.9
182.7	175.7	175.0	176.4	176.1	178.7	177.6	179.1	177.1	175.3
180.3	175.7	182.6	177.6	173.8	178.0	171.4	172.7	175.7	182.9
171.0	178.4	177.7	178.4	173.0	174.9	170.0	172.8	175.5	177.4

Ο έλεγχος

$$H_0 : \mu_A - \mu_B = 0 \quad - \quad H_1 : \mu_A - \mu_B \ne 0$$

εκτελείται στην $1^η$ Περίπτωση με το R ως ακολούθως:

```
> HA <- read.table("HEIGHTA.txt", header=T)
> HB <- read.table("HEIGHTB.txt", header=T)
> attach(HA);names(HA)
[1] "HEIGHTA"
> attach(HB);names(HB)
[1] "HEIGHTB"
> t.test(HEIGHTA, HEIGHTB, mu = 0, var.equal = TRUE,
+ alternative="two.sided", conf.level=0.95)

        Two Sample t-test

data:  HEIGHTA and HEIGHTB
t = -2.5545, df = 98, p-value = 0.01217
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.9347466 -0.3685867
sample estimates:
mean of x mean of y
 174.6683  176.3200
```

Πέραν της $p-value$ του ελέγχου δίνεται και το 95% διάστημα εμπιστοσύνης για το μέσο του πληθυσμού (λόγω του ορίσματος `conf.level=0.95`).

Επιβεβαίωση της *p-value*:

```
> n1 <- length(HEIGHTA); n2 <- length(HEIGHTB)
> var1 <- var(HEIGHTA); var2 <- var(HEIGHTB)
> mean1 <- mean(HEIGHTA); mean2 <- mean(HEIGHTB)
> s <- ((n1-1)*var1+(n2-1)*var2)/(n1+n2-2)
> t <- (mean1-mean2)/sqrt(s*(1/n1+1/n2))
> cat("Επιβεβαίωση: p-value =", 2*(1-pt(abs(t),n1+n2-2)), "\n")
Επιβεβαίωση: p-value = 0.01217217
```

Ο έλεγχος της $2^{ης}$ Περίπτωσης εκτελείται ως ακολούθως

```
> t.test(HEIGHTA, HEIGHTB, mu = 0, var.equal = FALSE,
+ alternative="two.sided", conf.level=0.95)

        Welch Two Sample t-test

data:  HEIGHTA and HEIGHTB
t = -2.4966, df = 77.041, p-value =0.01467
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.9689979 -0.3343354
sample estimates:
mean of x mean of y
 174.6683  176.3200
```

Επιβεβαίωση της *p-value*:

```
> dfnum <- (var1/n1+var2/n2)^2
> dfden <- var1^2/(n1^2*(n1-1))+var2^2/(n2^2*(n2-1))
> df <- dfnum/dfden
> t <- (mean1-mean2)/sqrt(var1/n1+var2/n2)
> cat("Επιβεβαίωση: p-value=", 2*(1-pt(abs(t),df)), "\n")
Επιβεβαίωση: p-value= 0.01467382
```

6.3.2 Έλεγχος για τη διαφορά δύο μέσων τιμών με δείγματα κατά ζεύγη

Έστω ανεξάρτητα ζεύγη (X_i, Y_i), $1 \le i \le n$, τα οποία ακολουθούν δισδιάστατη κανονική κατανομή με παραμέτρους $E(X) = \mu_1$, $E(Y) = \mu_2$, $Var(X) = \sigma_1^2$, $Var(Y) = \sigma_2^2$ και $\rho_{X,Y} = \rho$. Για τους ελέγχους

$$H_0 : \mu_D = \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_D = \mu_1 - \mu_2 < \delta_0,$$

$$H_0 : \mu_D = \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_D = \mu_1 - \mu_2 > \delta_0,$$

$$H_0 : \mu_D = \mu_1 - \mu_2 = \delta_0 \quad - \quad H_1 : \mu_D = \mu_1 - \mu_2 \ne \delta_0,$$

βασιζόμαστε στη στατιστική συνάρτηση ελέγχου

$$T = \frac{\overline{D} - \delta_0}{S_D / \sqrt{n}}$$

όπου $\overline{D}$ η δειγματική μέση τιμή και S_D η δειγματική τυπική απόκλιση των διαφορών $D_i = X_i - Y_i$, $1 \le i \le n$. Η T ακολουθεί κατανομή t_{n-1} όταν η H_0 είναι αληθής. Αν συμβολίσουμε με t την παρατηρούμενη τιμή της T, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : \mu_D < \delta_0$	$\{t : t < -t_{n-1;\,a}\}$	$P(T \le t)$
$H_1 : \mu_D > \delta_0$	$\{t : t > t_{n-1;\,a}\}$	$P(T \ge t)$
$H_1 : \mu_D \ne \delta_0$	$\{t : t < -t_{n-1;\,a/2}\} \cup \{t : t > t_{n-1;\,a/2}\}$	$2P(T \ge \lvert t \rvert) = 2(1 - P(T \le \lvert t \rvert))$

Ο παραπάνω έλεγχος μπορεί να χρησιμοποιηθεί χωρίς καμία περιοριστική υπόθεση για την κατανομή των παρατηρήσεων, αρκεί το μέγεθος του δείγματος να είναι μεγάλο.

Για τον έλεγχο της διαφοράς των μέσων τιμών μ_1 και μ_2 δύο πληθυσμών με δείγματα κατά ζεύγη χρησιμοποιείται πάλι η συνάρτηση `t.test` μαζί με το όρισμα `paired`. Η βασική σύνταξη της συνάρτησης `t.test` για τον έλεγχο της διαφοράς των μέσων τιμών δύο πληθυσμών με δείγματα κατά ζεύγη είναι η ακόλουθη

```
t.test(x, y, mu=δ0, paired=TRUE)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
mu: η τιμή της διαφοράς μ1-μ2 υπό τη μηδενική υπόθεση
```

Παράδειγμα 6.18 (t.test)

Τα δεδομένα του αρχείου WEIGHT.txt, περιέχουν βάρη σε *Kgr* 20 πειραματόζωων πριν (μεταβλητή WA, (X)) και μετά (μεταβλητή WB, (Y)) τη χορήγηση για μια εβδομάδα ενός σκευάσματος που

καταπολεμά την παχυσαρκία, και δίνονται στον ακόλουθο πίνακα

i	X_i	Y_i	i	X_i	Y_i
1	81.3	81.8	11	77.3	81.0
2	81.7	80.1	12	80.8	83.3
3	89.6	84.2	13	71.5	70.0
4	85.1	82.1	14	84.7	75.5
5	83.3	75.8	15	75.4	72.4
6	81.5	81.6	16	81.9	79.9
7	80.6	76.1	17	79.8	79.0
8	81.9	79.9	18	82.0	80.7
9	69.5	70.9	19	94.3	91.9
10	80.2	81.2	20	77.0	76.3

Για τον έλεγχο της ισότητας των μέσων τιμών των βαρών των πειραματόζωων πριν και μετά τη χορήγηση του σκευάσματος, δηλαδή για τον έλεγχο

$$H_0 : \mu_D = \mu_1 - \mu_2 = 0 \quad - \quad H_1 : \mu_D = \mu_1 - \mu_2 \neq 0$$

έχουμε τα ακόλουθα:

```
> W <- read.table("WEIGHT.txt", header=T)
> attach(W);names(W)
[1] "WA" "WB"
> t.test(WA, WB, mu = 0,  paired=TRUE, alternative="two.sided")

        Paired t-test

data:  WA and WB
t = 2.5302, df = 19, p-value = 0.02040
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3084184 3.2615816
sample estimates:
mean of the differences
                  1.785
```

Επιβεβαίωση της *p-value*:

```
> D <- WA-WB
> n <- length(WA)
> t <- mean(D)/(sd(D)/sqrt(n))
> cat("Επιβεβαίωση: p-value=", 2*pt(abs(t), df=n-1, lower.tail = FALSE))
Επιβεβαίωση: p-value= 0.02039696
```

Εναλλακτικά ο παραπάνω έλεγχος μπορεί να γίνει με `t.test` για το δείγμα των διαφορών $d_i = X_i - Y_i$, $i = 1, 2, ..., 20$.

```
> t.test(WA-WB, mu = 0, alternative="two.sided")

        One Sample t-test

data:  WA - WB
t = 2.5302, df = 19, p-value = 0.02040
alternative hypothesis: true mean is not equal to 0
```

6.3.3 Έλεγχος για το συντελεστή συσχέτισης δύο μεταβλητών

Ένας ιδιαίτερος έλεγχος που σχετίζεται με τον έλεγχο για δείγματα κατά ζεύγη είναι ο έλεγχος για το συντελεστή συσχέτισης ρ δύο τυχαίων μεταβλητών X και Y. Μας ενδιαφέρουν οι έλεγχοι υποθέσεων

$$H_0 : \rho = 0 \quad - \quad H_1 : \rho < 0 ,$$

$$H_0 : \rho = 0 \quad - \quad H_1 : \rho > 0 ,$$

$$H_0 : \rho = 0 \quad - \quad H_1 : \rho \neq 0 .$$

Οι εναλλακτικές υποθέσεις των τριών παραπάνω ελέγχων δηλώνουν ότι οι τυχαίες μεταβλητές X και Y είναι αρνητικά συσχετισμένες, θετικά συσχετισμένες, ασυσχέτιστες. Ειδικότερα στην περίπτωση που το ζεύγος (X, Y) ακολουθεί τη δισδιάστατη κανονική κατανομή ο τρίτος έλεγχος ισοδυναμεί με έλεγχο ανεξαρτησίας των X και Y. Η στατιστική συνάρτηση ελέγχου είναι η

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

όπου

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

(γραμμικός συντελεστής συσχέτισης του Pearson), η οποία ακολουθεί την κατανομή t_{n-2} όταν η H_0 είναι αληθής. Αν συμβολίσουμε με *t* την παρατηρούμενη τιμή της *T*, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$p-value$
$H_1 : \rho < 0$	$\{t : t < -t_{n-2;\,a}\}$	$P(T \le t)$
$H_1 : \rho > 0$	$\{t : t > t_{n-2;\,a}\}$	$P(T \ge t)$
$H_1 : \rho \neq 0$	$\{t : t < -t_{n-2;\,a/2}\} \bigcup \{t : t > t_{n-1;\,a/2}\}$	$2P(T \ge \lvert t \rvert) = 2(1 - P(T \le \lvert t \rvert))$

Ο παραπάνω έλεγχος γίνεται στο R με τη συνάρτηση `cor.test`. Η βασική σύνταξη της συνάρτησης `cor.test` δίνεται στο ακόλουθο πλαίσιο

```
cor.test(x, y, method=c("pearson", "kendall", "spearman"))
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
method: δηλώνει το συντελεστή συσχέτισης που θα χρησιμοποιηθεί
```

Παράδειγμα 6.19 (cor.test)

Αναφερόμενοι στα δεδομένα του Παραδείγματος 6.18, μας ενδιαφέρει ο΄έλεγχος υπόθεσης

$$H_0 : \rho = 0 \quad - \quad H_1 : \rho \neq 0 .$$

Χρησιμοποιώντας το R έχουμε τα ακόλουθα αποτελέσματα

```
> W <- read.table("WEIGHT.txt", header=T)
> attach(W);names(W)
[1] "WA" "WB"
> cor.test(WA, WB, method="pearson", alternative = "two.sided",
+ conf.level = 0.9)

        Pearson's product-moment correlation

data:  WA and WB
t = 6.1735, df = 18, p-value = 7.914e-06
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.6473194 0.9167935
sample estimates:
      cor
0.824146
```

Επιβεβαίωση της *p-value*:

```
> r <- cor(WA,WB)
> t <- r*sqrt(n-2)/sqrt(1-r^2)
> cat("Επιβεβαίωση: p-value=", 2*pt(abs(t), df=n-2, lower.tail = FALSE))
Επιβεβαίωση: p-value= 7.914368e-06
```

Επομένως οι δύο μεταβλητές δεν είναι ασυσχέτιστες (άρα ούτε ανεξάρτητες), οπότε δεν θα μπορούσαμε να χρησιμοποιήσουμε `t.test` για τον έλεγχο της διαφοράς των μέσων τιμών των δύο πληθυσμών (Παράγραφος 6.3.1). ■

6.3.4 Έλεγχος ισότητας των διακυμάνσεων δύο πληθυσμών

Έστω τα ανεξάρτητα δείγματα $X_1, X_2, ..., X_{n_1}$ και $Y_1, Y_2, ..., Y_{n_2}$ από κανονικούς πληθυσμούς $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, αντίστοιχα. Για τους ελέγχους

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \lambda_0 \quad - \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} < \lambda_0 ,$$

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \lambda_0 \quad - \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} > \lambda_0 ,$$

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \lambda_0 \quad - \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq \lambda_0 ,$$

χρησιμοποιείται η στατιστική συνάρτηση ελέγχου

$$F = \frac{1}{\lambda_0} \cdot \frac{S_1^2}{S_2^2}$$

η οποία ακολουθεί την κατανομή $F_{n_1-1,\, n_2-1}$ όταν η H_0 είναι αληθής. Αν συμβολίσουμε με *f* την πα-

ρατηρούμενη τιμή της F, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνεται στον ακόλουθο πίνακα

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} < \lambda_0$	$\{f : f < F_{n_1-1,\, n_2-1;\, 1-a}\}$	$P(F \le f)$
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > \lambda_0$	$\{f : f > F_{n_1-1,\, n_2-1;\, a}\}$	$P(F \ge f)$
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} \ne \lambda_0$	$\{f : f < F_{n_1-1,\, n_2-1;\, 1-a/2}\} \cup \{f : f > F_{n_1-1,\, n_2-1;\, a/2}\}$	$2\min\{P(F \le f), P(F \ge f)\}$

Στο R χρησιμοποιείται η συνάρτηση `var.test` για τον έλεγχο της ισότητας των διακυμάνσεων δύο κανονικών πληθυσμών. Η βασική σύνταξη της συνάρτησης `var.test` είναι η ακόλουθη

```
var.test(x, y, ratio=r)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
ratio: δηλώνει τον λόγο των διακυμάνσεων των πληθυσμών x και y με
       default τιμή r=1
```

Παράδειγμα 6.20 (var.test)

Έστω τα δύο ακόλουθα δείγματα ($n_1 = n_2 = 20$)

X: 135, 193, 98, 160, 62, 80, 75, 142, 132, 57, 213, 100, 75, 76, 93, 73, 133, 90, 151, 56

Y: 131, 123, 117, 271, 85, 126, 99, 195, 200, 54, 107, 121, 131, 61, 79, 191, 206, 105, 125, 57

Εκτελώντας τον έλεγχο

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad - \quad H_1 : \sigma_1^2 \ne \sigma_2^2$$

με το R παίρνουμε τα ακόλουθα:

```
> X <- c(135,193,98,160,62,80,75,142,132,57,213,100,
+75,76,93,73,133,90,151,56)
> Y <- c(131,123,117,271,85,126,99,195,200,54,107,121,
+ 131,61,79,191,206,105,125,57)
> var.test(X,Y, ratio=1)

        F test to compare two variances

data:  X and Y
F = 0.6376, num df = 19, denom df = 19, p-value = 0.3351
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2523864 1.6109706
sample estimates:
ratio of variances
         0.6376418
```

Επιβεβαίωση της *p-value*:

```
> f <- var(X)/var(Y)
> df1 <- length(X)-1; df2 <- length(Y)-1
> a <- 1-pf(f,df1,df2); b <- pf(f,df1,df2)
> cat("Επιβεβαίωση: p-value=", 2*min(a,b), "\n")
Επιβεβαίωση: p-value= 0.3350913
```

Από τα παραπάνω προκύπτει ότι η ισότητα των διακυμάνσεων των δύο πληθυσμών δεν μπορεί να απορριφθεί.

Από μελέτες έχει προκύψει ότι η στατιστική συνάρτηση ελέγχου F είναι ευαίσθητη στην υπόθεση της κανονικότητας των πληθυσμών. Όταν όμως το μέγεθος των δύο δειγμάτων είναι ίδιο τότε ο έλεγχος είναι λιγότερο ευαίσθητος σε αποκλίσεις από την υπόθεση της κανονικότητας των κατανομών των δύο πληθυσμών. Ωστόσο μια καλή τακτική σε περιπτώσεις που αμφιβάλλουμε για την κανονικότητα των δεδομένων είναι να προτιμήσουμε τον έλεγχο του Levene για την ισότητα των διακυμάνσεων. Ο έλεγχος του Levene μπορεί να εφαρμοστεί για τον έλεγχο της ισότητας των διακυμάνσεων k πληθυσμών και θα αναπτυχθεί επαρκώς σε επόμενη παράγραφο. Στο παρόν σημείο θα εφαρμοστεί χωρίς περαιτέρω λεπτομέρειες.

Στο R χρησιμοποιείται η συνάρτηση `levene.test` του πακέτου `lawstat` για τον έλεγχο της ισότητας των διακυμάνσεων δύο πληθυσμών (δείτε επίσης τη συνάρτηση `leveneTest` του πακέτου `car`). Η βασική σύνταξη της συνάρτησης `levene.test` είναι η ακόλουθη

```
levene.test(x, group, location=c("median", "mean", "trim.mean")
x: τα δεδομένα μας
group: μεταβλητή που δηλώνει σε ποιο πληθυσμό ανήκει το αντίστοιχο
    δεδομένο
location: δηλώνει το σημείο που θα χρησιμοποιηθεί για να υπολογι-
    στούν οι αποκλίσεις των δεδομένων μας από αυτό το σημείο.
```

Παράδειγμα 6.21 (levene.test)

Χρησιμοποιώντας τα δεδομένα του Παραδείγματος 6.20 παίρνουμε

```
> library(nortest)
------------------------------------------------------------------------
> lillie.test(X);shapiro.test(X)
        Lilliefors (Kolmogorov-Smirnov) normality test
D = 0.1846, p-value = 0.07256

        Shapiro-Wilk normality test
W = 0.9087, p-value = 0.06017
------------------------------------------------------------------------
> lillie.test(Y);shapiro.test(Y)
        Lilliefors (Kolmogorov-Smirnov) normality test
D = 0.2374, p-value = 0.004395

        Shapiro-Wilk normality test
W = 0.9132, p-value = 0.07343
```

Με τα παραπάνω αποτελέσματα δεν μπορούμε να υποστηρίξουμε με βεβαιότητα ότι οι δύο πληθυσμοί είναι κανονικοί. Επομένως η χρήση του ελέγχου Levene θεωρείται επιβεβλημένη.

```
> library(lawstat)
> V <- c(X,Y)
> G <- c(rep(1,times=length(X)),rep(2,times=length(Y)))
-------------------------------------------------------------------------
> levene.test(V, G, location="mean")

        classical Levene's test based on the absolute deviations from the
mean ( none not applied because the location is not set to median )

data:  V
Test Statistic = 0.1616, p-value = 0.69
-------------------------------------------------------------------------
> levene.test(V, G, location="median")

        modified robust Brown-Forsythe Levene-type test based on the ab-
solute deviations from the median

data:  V
Test Statistic = 0.18, p-value = 0.6738
```

Από τα παραπάνω προκύπτει (όπως και με το `var.test`) ότι η ισότητα των διακυμάνσεων των δύο πληθυσμών δεν μπορεί να απορριφθεί. ■

6.3.5 Έλεγχοι για αναλογίες επιτυχιών

6.3.5.1 Ακριβής έλεγχος του Fisher για την ισότητα δύο αναλογιών επιτυχιών

Έστω τα ανεξάρτητα δείγματα $X_1, X_2, ..., X_{n_1}$ και $Y_1, Y_2, ..., Y_{n_2}$ από τους πληθυσμούς $B(1, p_1)$ και $B(1, p_2)$, αντίστοιχα. Οι στατιστικές συναρτήσεις $X = \sum_{i=1}^{n_1} X_i$ και $Y = \sum_{i=1}^{n_2} Y_i$ δηλώνουν τον αριθμό των επιτυχιών στο πρώτο και στο δεύτερο δείγμα αντίστοιχα. Μας ενδιαφέρουν οι έλεγχοι

$$H_0 : p_1 = p_2 \quad - \quad H_1 : p_1 < p_2 \quad (H_0 : p_1 - p_2 = 0 \quad - \quad H_1 : p_1 - p_2 < 0)$$

$$H_0 : p_1 = p_2 \quad - \quad H_1 : p_1 > p_2 \quad (H_0 : p_1 - p_2 = 0 \quad - \quad H_1 : p_1 - p_2 > 0)$$

$$H_0 : p_1 = p_2 \quad - \quad H_1 : p_1 \neq p_2 \quad (H_0 : p_1 - p_2 = 0 \quad - \quad H_1 : p_1 - p_2 \neq 0)$$

Αν x και y είναι οι παρατηρούμενες τιμές των X και Y τότε έχουμε τον ακόλουθο 2×2 πίνακα συνάφειας

	Επιτυχίες	Αποτυχίες	Σύνολο
Δείγμα 1	x	$n_1 - x$	n_1
Δείγμα 2	y	$n_2 - y$	n_2
Σύνολο	k	$N - k$	N

Στον ακριβή έλεγχο του Fisher υποθέτουμε ότι τα περιθώρια αθροίσματα του παραπάνω πίνακα είναι σταθερά. Τότε ισχύει ότι

$$P(X=i \mid X+Y=k)=P(U_1=i)=\frac{\binom{n_1}{i}\binom{n_2}{k-i}}{\binom{N}{k}},\quad i=\max\{0,k-n_2\},\ldots,\min\{n_1,k\},$$

όπου η τυχαία μεταβλητή U_1 ακολουθεί την υπεργεωμετρική κατανομή με παραμέτρους n_1, n_2, και k ($Hyper(n_1,n_2,k)$). Σημειώνουμε ότι $E(U_1)=kn_1/(n_1+n_2)$. Ο ακριβής έλεγχος του Fisher κρίνει αν ο αριθμός των επιτυχιών X (στατιστική συνάρτηση ελέγχου) είναι σημαντικά μικρός ή σημαντικά μεγάλος. Για παράδειγμα, στον έλεγχο

$$H_0: p_1=p_2 \quad - \quad H_1: p_1<p_2$$

η μηδενική υπόθεση απορρίπτεται αν ο αριθμός των επιτυχιών X είναι αρκετά μικρός.

Ο υπολογισμός της $p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	***p - value***
$H_1: p_1<p_2$	$P(X\le x)=\sum_{i=\max\{0,k-n_2\}}^{x}\binom{n_1}{i}\binom{n_2}{k-i}\binom{N}{k}^{-1}$
$H_1: p_1>p_2$	$P(X\ge x)=\sum_{i=x}^{\min\{n_1,k\}}\binom{n_1}{i}\binom{n_2}{k-i}\binom{N}{k}^{-1}$
$H_1: p_1\ne p_2$	$\sum_{i=\max\{0,k-n_2\}}^{\min\{n_1,k\}} I[P(X=i)\le P(X=x)]\binom{n_1}{i}\binom{n_2}{k-i}\binom{N}{k}^{-1}$

Ο ακριβής έλεγχο του Fisher χρησιμοποιείται παραδοσιακά όταν τα μεγέθη των δύο δειγμάτων είναι μικρά. Όταν τα μεγέθη των δύο δειγμάτων είναι μεγάλα τότε μπορεί να χρησιμοποιηθεί ένας έλεγχος χ^2 (δείτε Παράγραφο 6.4.2) που αποδίδει μια προσεγγιστική $p-value$. Ωστόσο αν οι μέσες τιμές $E(U_1)=kn_1/(n_1+n_2)$ και $E(U_2)=kn_2/(n_1+n_2)$ είναι αμφότερες μικρότερες του 10 τότε συνιστάται αποκλειστικά η χρήση του ελέγχου του Fisher.

Ο ακριβής έλεγχος του Fisher γίνεται στο R με τη συνάρτηση `fisher.test` που έχει την ακόλουθη βασική σύνταξη

```
fisher.test(m, or)
m: ένας 2x2 πίνακας συνάφειας
or: η τιμή του odds ratio [p1/(1- p1)]/[p2/(1- p2)]
```

Παράδειγμα 6.22 (έλεγχος ισότητας δύο ποσοστών – ακριβής έλεγχος του Fisher)

Θέλουμε να ελέγξουμε αν το ποσοστό των γυναικών που κάνουν δίαιτα είναι μεγαλύτερο από το

αντίστοιχο ποσοστό των ανδρών. Επιλέγουμε τυχαία 14 γυναίκες και 10 άνδρες και τα αποτελέσματα δίνονται στον ακόλουθο πίνακα.

		Δίαιτα		
		Ναι	Όχι	Σύνολο
Φύλο	Γυναίκες	10	4	14
	Άνδρες	2	8	10
	Σύνολο	12	12	24

Ο έλεγχος υπόθεσης που μας ενδιαφέρει είναι ο ακόλουθος

$$H_0 : p_1 = p_2 \quad - \quad H_1 : p_1 > p_2$$

ή ισοδύναμα

$$H_0 : \frac{p_1(1-p_2)}{p_2(1-p_1)} = 1 \quad - \quad H_1 : \frac{p_1(1-p_2)}{p_2(1-p_1)} > 1$$

όπου p_1 (p_2) είναι το ποσοστό των γυναικών (ανδρών) που κάνουν δίαιτα.

```
> m <- matrix(c(10,2,4,8), nrow=2)
> m
     [,1] [,2]
[1,]   10    4
[2,]    2    8
> fisher.test(m, alternative="greater", or=1)

        Fisher's Exact Test for Count Data

data:  m
p-value = 0.01804
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.435748      Inf
sample estimates:
odds ratio
  8.913675
```

Επιβεβαίωση της *p-value*:

```
> pvalue <- sum(dhyper(10:12, 14, 10, 12))
> cat("Επιβεβαίωση: p-value=", pvalue, "\n")
Επιβεβαίωση: p-value= 0.01803742
```

Αξίζει να σημειώσουμε ότι $E(U_1) = 7$ και $E(U_2) = 5$. ■

Ο ακριβής έλεγχος του Fisher μπορεί να χρησιμοποιηθεί και ως έλεγχος ανεξαρτησίας ή ομογένειας σε πίνακες συνάφειας 2×2 (δείτε Παράγραφο 6.4.2).

6.3.5.2 Προσεγγιστικός έλεγχος για την ισότητα δύο αναλογιών επιτυχιών

Οι έλεγχοι της προηγούμενης παραγράφου, όταν τα μεγέθη των δειγμάτων n_1, n_2 είναι μεγάλα, μπορούν να διεξαχθούν προσεγγιστικά βασιζόμενοι στη στατιστική συνάρτηση ελέγχου

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

όπου

$$\hat{p}_1 = \frac{X}{n_1}, \quad \hat{p}_2 = \frac{Y}{n_2}, \quad \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}.$$

Η Z έχει προσεγγιστικά την τυπική κανονική κατανομή (για μεγάλες τιμές των n_1, n_2) όταν η H_0 είναι αληθής. Αν συμβολίσουμε με z την παρατηρούμενη τιμή της Z, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ δίνονται στον ακόλουθο πίνακα

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	***p – value***
$H_1 : p_1 < p_2$	$\{z : z < -z_a\}$	$P(Z \le z) = \Phi(z)$
$H_1 : p_1 > p_2$	$\{z : z > z_a\}$	$P(Z \ge z) = 1 - \Phi(z)$
$H_1 : p_1 \ne p_2$	$\{z : z < -z_{a/2}\} \bigcup \{z : z > z_{a/2}\}$	$2P(Z \ge \lvert z \rvert) = 2(1 - \Phi \lvert z \rvert)$

Όταν $|\hat{p}_1 - \hat{p}_2| > (1/2)[(1/n_1)+(1/n_2)]$ χρησιμοποιείται συνήθως ένα είδος διόρθωσης συνέχειας στην παρατηρούμενη τιμή z της Z, η οποία δίνεται στον ακόλουθο πίνακα

Συνθήκη	***z***	**Συνθήκη**	***z***
$\hat{p}_1 - \hat{p}_2 > 0$	$\dfrac{\hat{p}_1 - \hat{p}_2 - \frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	$\hat{p}_1 - \hat{p}_2 < 0$	$\dfrac{\hat{p}_1 - \hat{p}_2 + \frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

Ο προσεγγιστικός έλεγχος για την ισότητα δύο αναλογιών με ανεξάρτητα δείγματα γίνεται με τη βοήθεια της συνάρτησης `prop.test`. Η βασική σύνταξη της συνάρτησης `prop.test` είναι η ακόλουθη:

```
prop.test(x, n, correct=TRUE or FALSE)
x: διάνυσμα με τον αριθμό των επιτυχιών
n: διάνυσμα με τον αριθμό των δοκιμών
correct: υπολογισμός p-value με διόρθωση (TRUE) ή χωρίς διόρθωση
    (FALSE) συνέχειας
```

Παράδειγμα 6.23 (έλεγχος ισότητας δύο ποσοστών – προσεγγιστικός έλεγχος)

Σε δύο ανεξάρτητα τυχαία δείγματα μεγέθους 1000 το καθένα, ένα δείγμα ανδρών και ένα δείγμα γυναικών, βρέθηκαν αντίστοιχα 100 και 120 άτομα, που υποστηρίζουν κάποιον πολιτικό. Για να διαπιστωθεί αν το ποσοστό p_1 των ανδρών και το ποσοστό p_2 των γυναικών που υποστηρίζουν το

συγκεκριμένο πολιτικό διαφέρουν θα εκτελεστεί ο έλεγχος της υπόθεσης

$$H_0 : p_1 - p_2 = 0 \quad - \quad H_1 : p_1 - p_2 \neq 0 .$$

Χρησιμοποιώντας το R παίρνουμε

```
> prop.test(c(100,120), c(1000,1000), alternative="two.sided",
+ conf.level=0.95, correct=FALSE)

2-sample test for equality of proportions without continuity correction

data:  c(100, 120) out of c(1000, 1000)
X-squared = 2.0429, df = 1, p-value = 0.1529
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.047411482  0.007411482
sample estimates:
prop 1 prop 2
  0.10   0.12
```

Επιβεβαίωση της *p-value*:

```
> s1 <- 100; n1 <- 1000; p1 <- s1/n1; s2 <- 120; n2 <- 1000; p2 <- s2/n2
> p <- (n1*p1+n2*p2)/(n1+n2)
> z <- (p1-p2)/sqrt(p*(1-p)*((1/n1)+(1/n2)))
> cat("Επιβεβαίωση: p-value=", 2*(1-pnorm(abs(z))), "\n")
Επιβεβαίωση: p-value= 0.1529178
```

Παρατηρούμε ότι το R δίνει ως τιμή της στατιστικής συνάρτησης ελέγχου το z^2 (είναι γνωστό ότι η Z^2, όταν η μηδενική υπόθεση είναι αληθής, ακολουθεί την κατανομή χ_1^2).

Επειδή $0.02 = |\hat{p}_1 - \hat{p}_2| > (1/2)[(1/n_1)+(1/n_2)] = 0.001$, κρίνεται απαραίτητο να χρησιμοποιηθεί διόρθωση συνέχειας. Έτσι

```
> prop.test(c(100,120), c(1000,1000), alternative="two.sided",
+ conf.level=0.95, correct=TRUE)

2-sample test for equality of proportions with continuity correction

data:  c(100, 120) out of c(1000, 1000)
X-squared = 1.8437, df = 1, p-value = 0.1745
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.048411482  0.008411482
sample estimates:
prop 1 prop 2
  0.10   0.12
```

Επειδή $E(U_1) = E(U_2) = 110 > 10$, ο προσεγγιστικός έλεγχος και ο έλεγχος του Fisher θα πρέπει να έχουν σχεδόν ίδιες *p-value*. Πράγματι

```
> m <- matrix(c(100,120,900,880), nrow=2)
> fisher.test(m, alternative="two.sided", or=1)

        Fisher's Exact Test for Count Data

data:  m
```

```
p-value = 0.1744
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.608374 1.089736
sample estimates:
odds ratio
 0.8149004
```

Ο προσεγγιστικός έλεγχος

$$H_0 : p_1 = p_2 \quad - \quad H_1 : p_1 \neq p_2$$

που παρουσιάσαμε στην παρούσα παράγραφο είναι ισοδύναμος με τον έλεγχο ομογένειας σε πίνακες συνάφειας 2×2 (δείτε Παράγραφο 6.10.3.2). Απλά εδώ αναφέρουμε ότι χρησιμοποιώντας τους συμβολισμούς και τις αντιστοιχίσεις του παρακάτω πίνακα

	Επιτυχίες	Αποτυχίες	Σύνολο
Δείγμα 1	$x\ (n_{11})$	$n_1 - x\ (n_{12})$	$n_1\ (n_{1\bullet})$
Δείγμα 2	$y\ (n_{21})$	$n_2 - y\ (n_{22})$	$n_2\ (n_{2\bullet})$
Σύνολο	$k\ (n_{\bullet 1})$	$N - k\ (n_{\bullet 2})$	$N\ (n_{\bullet\bullet})$

έχουμε ότι

$$Z^2 = U = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(n_{ij} - \dfrac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\dfrac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}} \sim \chi_1^2 .$$

Για περισσότερες λεπτομέρειες δείτε Παράγραφο 6.4.2.

6.3.6 Q-Q διάγραμμα για δύο δείγματα

Στην Παράγραφο 6.2.7.3 παρουσιάσαμε το Q-Q διάγραμμα το οποίο μας επιτρέπει να αντιληφθούμε αν ένα τυχαίο δείγμα προέρχεται από πληθυσμό με συνάρτηση κατανομής $F(x)$. Υπάρχει ένα αντίστοιχο διάγραμμα που χρησιμοποιείται όταν υπάρχουν δύο ανεξάρτητα δείγματα, έστω το $X_1,\ X_2,\ ....,\ X_n$ και τ $Y_1,\ Y_2,\ ....,\ Y_m$, με το οποίο μπορούμε να αντιληφθούμε αν οι κατανομές των δύο συνόλων δεδομένων είναι ίδιες ή διαφέρουν μόνο ως προς θέση/κλίμακα.

Όταν τα δύο σύνολα των δεδομένων έχουν το ίδιο μέγεθος, το διάγραμμα αποτελείται από τα ζεύγη $(x_{(i)}, y_{(i)})$, $i = 1,2,...,n.$ Όταν τα μεγέθη είναι άνισα, έστω $n > m$, τότε το διάγραμμα αποτελείται από τα ζεύγη $(x_{p_i}, y_{(i)})$, $i = 1,2,...,m,$ όπου

$$p_i = (i-a)/(n+1-2a), \quad a = \begin{cases} 3/8, & n \le 10 \\ 1/2, & n > 10 \end{cases}$$

(το σημείο x_{p_i} αποτελεί το (κάτω) p_i – ποσοστιαίο σημείο του πρώτου δείγματος).

Αν τα σημεία του διαγράμματος πέφτουν σε μια νοητή ευθεία τότε τα δύο σύνολα δεδομένων προέρχονται από την ίδια οικογένεια κατανομών. Αν η νοητή ευθεία συμπίπτει με την $y = x$ οι δύο κατανομές είναι ίδιες, ενώ αν είναι παράλληλη προς την $y = x$ διαφέρουν μόνο ως προς θέση.

Για τη δημιουργία του παραπάνω Q-Q διαγράμματος χρησιμοποιούμε τη συνάρτηση `qqplot` με την ακόλουθη απλή σύνταξη

```
qqplot(x, y)
x: το πρώτο δείγμα
y: το δεύτερο δείγμα
```

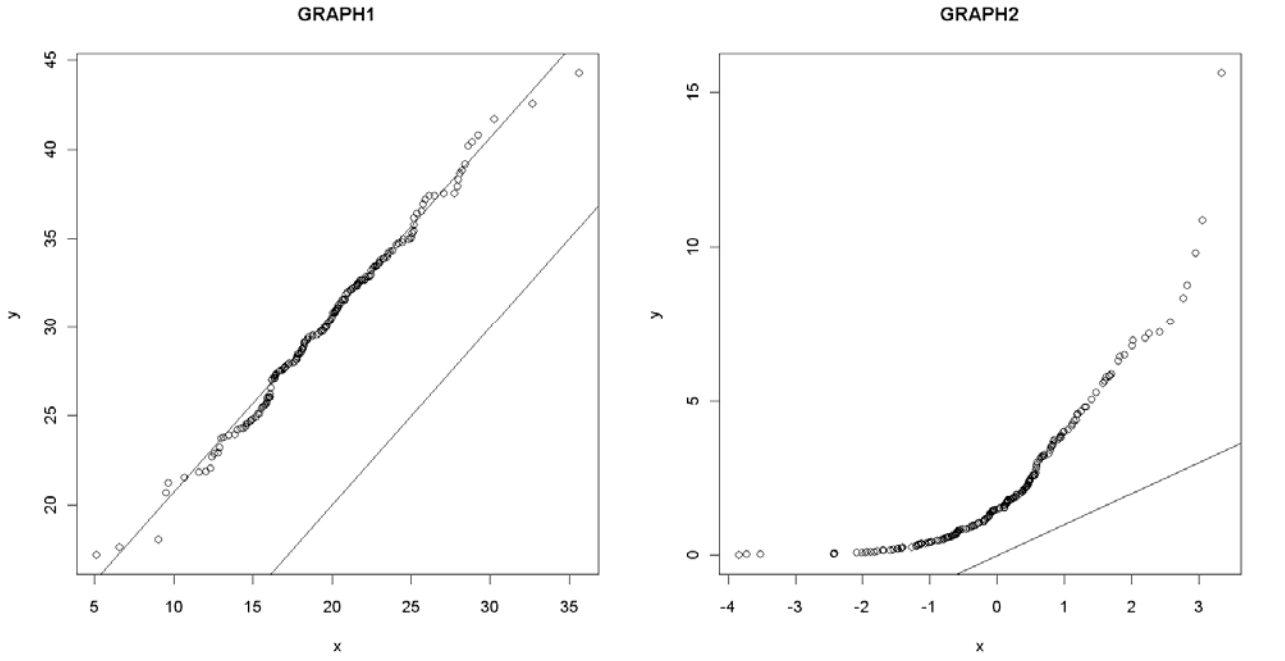
Παράδειγμα 6.24 (qqplot με δύο δείγματα)

Αρχικά κατασκευάζουμε δύο ανεξάρτητα τυχαία δείγματα μεγέθους 200 και 180 το καθένα, από τις κατανομές $N(20, 25)$ και $N(30, 25)$, αντίστοιχα. Παρατηρούμε ότι οι δύο πληθυσμοί διαφέρουν μόνο ως προς τη θέση και επομένως τα σημεία στο Q-Q διάγραμμα θα πρέπει να βρίσκονται πάνω σε μια ευθεία παράλληλη με την ευθεία $y = x$ (δείτε GRAPH1).

Στη συνέχεια δίνουμε το Q-Q διάγραμμα δύο δειγμάτων που το ένα προέρχεται από την κατανομή t_5 και το άλλο από την χ^2_2. Τα σημεία του διαγράμματος δεν πρέπει να βρίσκονται γύρω από μια ευθεία (δείτε GRAPH2).

```
> par(mfrow=c(1,2))
> x <- rnorm(200, mean=20, sd=5); y <- rnorm(180, mean=30, sd=5)
> qqplot(x,y, main="GRAPH1")
> abline(0,1)
> a1 <- quantile(x, 0.25); b1 <- quantile(y, 0.25)
> a2 <- quantile(x, 0.75); b2 <- quantile(y, 0.75)
```

```
> abline(b1-a1*((b2-b1)/(a2-a1)),(b2-b1)/(a2-a1))
-------------------------------------------------------------------------------
> x <- rt(200, df=5); y <- rchisq(300, df=2)
> qqplot(x,y,main="GRAPH2")
> abline(0,1)
```



6.3.7 Έλεγχος Kolmogorov-Smirnov

Ο έλεγχος Kolmogorov-Smirnov για δύο δείγματα χρησιμοποιείται για να ελέγξουμε αν δύο ανεξάρτητα τυχαία δείγματα $X_1, X_2, ..., X_n$ (από τον πληθυσμό X) και $Y_1, Y_2, ..., Y_m$ (από τον πληθυσμό Y) προέρχονται από την ίδια συνεχή κατανομή. Μας ενδιαφέρει ο έλεγχος

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \neq F_Y(x)$$

(η H_0 ισχύει για όλα τα x ενώ η H_1 ισχύει για τουλάχιστον ένα x) όπου $F_X(x)$ και $F_Y(x)$ οι συναρτήσεις κατανομής των δύο πληθυσμών X και Y. Επίσης μας ενδιαφέρουν οι μονόπλευροι έλεγχοι

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \geq F_Y(x)$$

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \leq F_Y(x)$$

(η H_0 ισχύει για όλα τα x ενώ στις H_1 ισχύει γνήσια ανισότητα για ένα τουλάχιστον x [†]).

Ενώ η μηδενική υπόθεση είναι σαφής, οι παραπάνω εναλλακτικές υποθέσεις είναι αρκετά γενικές. Στην πράξη, γίνονται άλλες πιο «συγκεκριμένες» εναλλακτικές υποθέσεις. Σύμφωνα με το **μοντέλο**

[†] Αν $F_X(x) \geq F_Y(x)$ με αυστηρή ανισότητα για κάποιο x τότε λέμε ότι η τυχαία μεταβλητή Y είναι στοχαστικά μεγαλύτερη από την τυχαία μεταβλητή X (συμβολισμός $Y \overset{ST}{>} X$).

θέσης (location model) ή αλλιώς μοντέλο μετατόπισης (shift model) υποθέτουμε ότι οι πληθυσμοί X και Y είναι ίδιοι εκτός ίσως από μια πιθανή διαφοροποίηση στη θέση τους, η οποία εκφράζεται από μια μετατόπιση θ, δηλαδή ότι οι τυχαίες μεταβλητές $X+\theta$ και Y έχουν την ίδια κατανομή (ή ότι η X και η $Y-\theta$ έχουν την ίδια κατανομή). Αν $Y = X+\theta$, τότε

$$F_Y(x) = P(Y \le x) = P(X+\theta \le x) = F_X(x-\theta), \quad x \in R,\ \theta \ne 0 .$$

Σύμφωνα με το μοντέλο θεσης, ο έλεγχος

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \ne F_Y(x)$$

μπορεί να γραφεί ως

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_Y(x) \ne F_X(x-\theta) \ \text{ για κάποιο } \theta \ne 0$$

ή ισοδύναμα

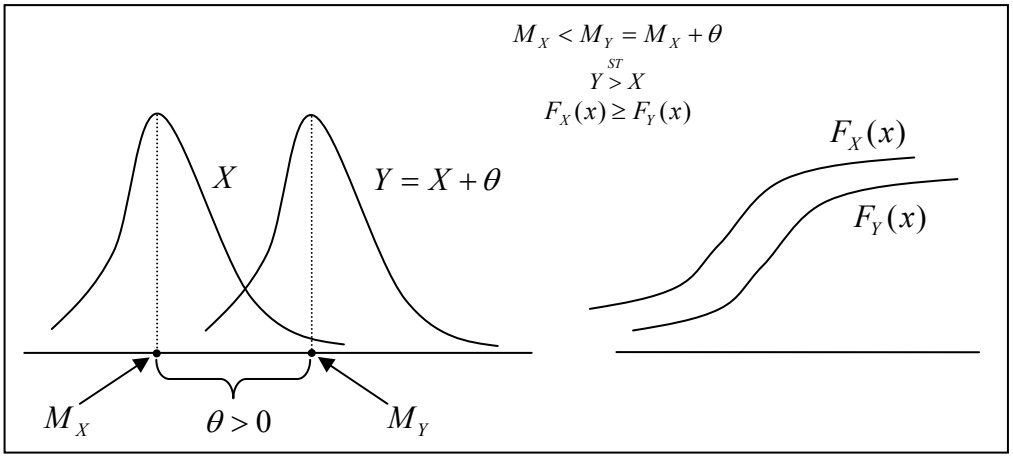
$$H_0 : \theta = 0 \quad - \quad H_1 : \theta \ne 0 .$$

Επίσης έχουμε τις κάτωθι ισοδύαμες εκφράσεις υπό το μοντέλο θέσης

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \ge F_Y(x) \quad \text{ή ισοδύναμα} \quad H_0 : \theta = 0 \quad - \quad H_1 : \theta > 0 ,$$

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \le F_Y(x) \quad \text{ή ισοδύναμα} \quad H_0 : \theta = 0 \quad - \quad H_1 : \theta < 0 .$$

Ενδεικτικό είναι το ακόλουθο σχήμα.



Σύμφωνα με το μοντέλο θέσης οι δύο πληθυσμοί έχουν την ίδια μορφή και την ίδια διακύμανση, ενώ η παράμετρος θ είναι συνήθως ίση είτε με τη διαφορά των μέσων $\mu_Y - \mu_X$, είτε με τη διαφορά των διαμέσων $M_Y - M_X$, είτε γενικότερα με τη διαφορά δύο αντίστοιχων παραμέτρων θέσης ή ακόμη και ποσοστιαίων σημείων.

Επίσης υπάρχει και το μοντέλο κλίμακας (scale model) και το μεικτό μοντέλο (location-scale model)[†]

[†] Σύμφωνα με το μοντέλο κλίμακας η X/θ $(\theta \ne 1)$ και η Y έχουν την ίδια κατανομή. Τότε, για παράδειγμα $F_Y(x) = P(Y \le x) = P(X \le \theta x) = F_X(\theta x), \quad x \in R,\ \theta > 0,\ \theta \ne 1 .$

Για τον έλεγχο

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \neq F_Y(x)$$

χρησιμοποιείται η στατιστική συνάρτηση ελέγχου $D_{n,m}$ που δίνεται από τη σχέση

$$D_{n,m} = \sup_{-\infty<x<\infty} \{|S_n(x) - S_m(x)|\}$$

όπου $S_n(x)$ ($S_m(x)$) η εμπειρική συνάρτηση κατανομής του πρώτου (δεύτερου) δείγματος. Προφανώς μεγάλες τιμές της $D_{n,m}$ προσφέρουν ενδείξεις για την ισχύ της εναλλακτικής υπόθεσης. Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$$\{d_{n,m} : \ d_{n,m} > D_{n,m;\,a}\}$$

όπου $d_{n,m}$ είναι η παρατηρούμενη τιμή της $D_{n,m}$, και $D_{n,m;\,a}$ το άνω a ποσοστιαίο σημείο της κατανομής της $D_{n,m}$. Η $p-value$ του ελέγχου είναι ίση με $P(D_{n,m} > d_{n,m})$. Επειδή η ακριβής κατανομή της $D_{n,m}$ είναι δύσκολο να υπολογιστεί, και για να διευκολυνθεί η εκτέλεση του ελέγχου, έχουν κατασκευαστεί ειδικοί πίνακες με τα ποσοστιαία σημεία της $D_{n,m}$. Ωστόσο, μπορεί να αποδειχθεί ότι

$$\lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \le d\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

οπότε μια προσεγγιστική τιμή της $p-value$ του ελέγχου για μεγάλα δείγματα είναι η

$$p-value = P(D_{n,m} > d_{n,m}) = 1 - P(D_{n,m} \le d_{n,m}) \cong 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2\frac{nm}{n+m} i^2 d_{n,m}^2}.$$

Με ανάλογο τρόπο, όπως και στην περίπτωση του ελέγχου Kolmogorov-Smirnov για ένα δείγμα, μπορούν να αναπτυχθούν και οι μονόπλευροι έλεγχοι. Για παράδειγμα, ο έλεγχος

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \ge F_Y(x)$$

εκτελείται με τη βοήθεια της στατιστικής συνάρτησης

$$D_{n,m}^{+} = \sup_{-\infty<x<\infty} [S_n(x) - S_m(x)]$$

και η κρίσιμη περιοχή δίνεται από τη σχέση $\{d_{n,m}^{+} : \ d_{n,m}^{+} \ge D_{n,m;\,a}^{+}\}$.

Για την εκτέλεση του ελέγχου Kolmogorov-Smirnov για δύο δείγματα χρησιμοποιείται η συνάρτηση `ks.test`. Η βασική σύνταξη της συνάρτησης `ks.test` για δύο δείγματα είναι η ακόλουθη

```
ks.test(x, y)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
```

Σύμφωνα με το μεικτό μοντέλο η $(X - \mu_X)/\theta$ και η $Y - \mu_Y$ έχουν την ίδια κατανομή (ομοίως και για M_X, M_Y στη θέση των μ_X, μ_Y), οπότε $P(Y - \mu_Y \le x) = P(X - \mu_X \le \theta x)$.

Παράδειγμα 6.25 (ks.test – δίπλευρη εναλλακτική)

Το αρχείο wings.txt περιέχει τα μήκη των φτερών από έντομα που βρέθηκαν σε δύο διαφορετικές γεωγραφικές περιοχές την Α ($n = 50$) και τη Β ($m = 70$). Τα δεδομένα παρουσιάζονται στον ακόλουθο πίνακα

Περιοχή Α					Περιοχή Β						
29.3114	22.73126	26.11796	32.91597	27.90548	29.81984	29.38996	16.891	29.86668	16.06792	29.3669	19.50692
25.96292	23.48272	29.06467	24.23003	22.82924	25.58656	17.82669	28.58179	30.05387	30.2029	26.98399	17.15765
24.47853	25.81436	22.06382	23.11119	24.0134	25.08861	27.87755	24.75652	23.35198	26.30017	24.55058	27.53538
22.94904	23.61457	22.88138	19.48177	19.48367	29.47527	23.40487	27.9327	21.43363	30.51081	24.88922	29.07388
22.25191	24.72731	23.44479	29.05983	22.19061	29.4917	32.02936	28.7529	22.10865	15.18989	18.89952	30.81347
25.20959	28.36228	21.18421	22.77284	23.63771	23.38806	28.57462	25.62206	28.06292	27.29669	26.95935	21.8908
19.33376	24.51804	20.47674	23.14654	26.165	24.27135	32.44551	23.56552	22.74691	30.14747	34.19588	14.76077
25.78789	16.98068	23.77729	26.00667	28.8833	20.9552	25.92111	23.08194	21.86922	23.99693	24.66259	21.59737
22.56246	28.37925	20.26934	22.45082	28.86669	29.63321	31.36365	22.88308	20.94718	23.05739	27.19062	20.59379
25.52483	24.2059	19.18757	22.07384	21.99278	25.35347	27.4341	17.25769	26.0353	28.78991	23.2039	25.92797

Θέλουμε να ελέγξουμε αν οι κατανομές των μηκών των φτερών των εντόμων στις δύο περιοχές είναι ίδιες. Συμβολίζοντας με $F_X(x)$ ($F_Y(x)$) τη συνάρτηση κατανομής των μηκών των φτερών στην περιοχή Α (Β) θα εκτελέσουμε με το R τον έλεγχο

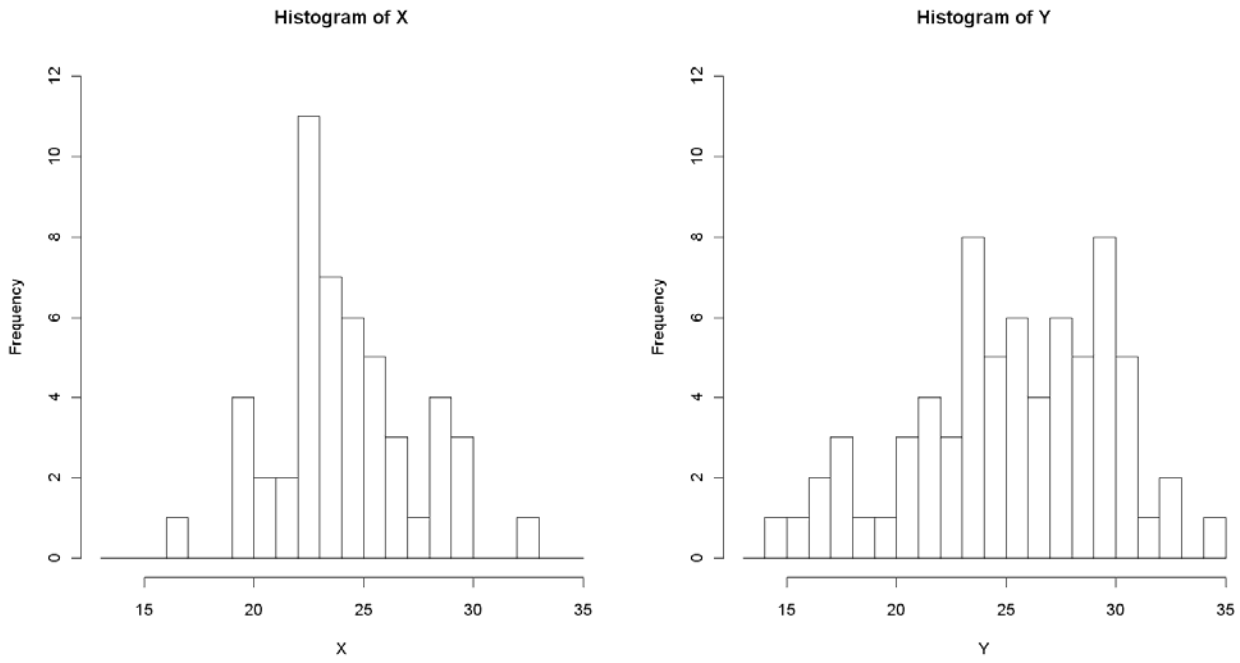
$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \neq F_Y(x).$$

```
> wings <- read.table("wings.txt", header=TRUE)
> names(wings);attach(wings)
[1] "size"     "location"
> wings
        size location
1   29.31140        A
2   25.96292        A
⋮       ⋮           ⋮
119 20.59379        B
120 25.92797        B
> X<-size[location=="A"]; Y<-size[location=="B"]
-------------------------------------------------------------------------
> ks.test(X, Y, exact=TRUE)

        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.2629, p-value = 0.02911
alternative hypothesis: two-sided
-------------------------------------------------------------------------
> ks.test(X, Y, exact=FALSE)

        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.2629, p-value = 0.03553
alternative hypothesis: two-sided
```

Από τα παραπάνω προκύπτει ότι σε επίπεδο σημαντικότητας 0.05, είτε με τον ακριβή έλεγχο είτε με τον προσεγγιστικό έλεγχο, οι δύο κατανομές διαφέρουν (αν δεν υπήρχε το όρισμα `exact` τότε

το R θα έδινε ως αποτέλεσμα την ακριβή τιμή της *p-value* αν $nm < 10000$, ειδάλλως θα έδινε την προσεγγιστική τιμή της). Αποφασίζουμε να κατασκευάσουμε τα ιστογράμματα των δύο συνόλων δεδομένων για να αντιληφθούμε που διαφέρουν οι δύο κατανομές.

```
> par(mfrow=c(1,2))
> class <- seq(13,35,1)
> hist(X, breaks=class, xlim=c(13,35), ylim=c(0,12))
> hist(Y, breaks=class, xlim=c(13,35),ylim=c(0,12))
```



Παρατηρούμε ότι η διασπορά των παρατηρήσεων φαίνεται να είναι μεγαλύτερη στην περιοχή B, ενώ οι μέσες τιμές τους δεν φαίνεται να διαφέρουν. Πράγματι, σε επίπεδο σημαντικότητας 0.05, από το παρακάτω πλαίσιο προκύπτει ότι οι μέσες τιμές δεν διαφέρουν, σε αντίθεση με τις διακυμάνσεις.

```
> t.test(X,Y,var.equal = FALSE)

        Welch Two Sample t-test

data:  X and Y
t = -1.6073, df = 117.996, p-value = 0.1107
alternative hypothesis: true difference in means is not equal to 0
------------------------------------------------------------------------
> var.test(X,Y,ratio=1)

        F test to compare two variances

data:  X and Y
F = 0.5014, num df = 49, denom df = 69, p-value = 0.01192
alternative hypothesis: true ratio of variances is not equal to 1
```

Παράδειγμα 6.26 (ks.test – μονόπλευρη εναλλακτική)

Για τον έλεγχο της υπόθεσης

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \geq F_Y(x)$$

δίνουμε το ακόλουθο παράδειγμα το οποίο έχει κατασκευαστεί με τέτοιο τρόπο ώστε να είναι αληθής η εναλλακτική υπόθεση.

```
> x <- rnorm(150, mean=50, sd=4)
> y <- rnorm(110, mean=51.7, sd=4)
> ks.test(x,y, alternative="greater")

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D^+ = 0.2224, p-value = 0.001875
alternative hypothesis: the CDF of x lies above that of y
```

6.3.8 Προσημικός έλεγχος με δείγματα κατά ζεύγη (sign test)

Έστω ένα τυχαίο δείγμα ζευγαρωτών παρατηρήσεων (X_i, Y_i), $1 \leq i \leq n$. Υποθέτουμε ότι οι διαφορές $D_i = X_i - Y_i$ είναι ανεξάρτητες τυχαίες μεταβλητές προερχόμενες από ένα συνεχή πληθυσμό D με διάμεσο M. Μας ενδιαφέρουν οι έλεγχοι

$$H_0 : M = 0 \quad - \quad H_1 : M \neq 0,$$

$$H_0 : M = 0 \quad - \quad H_1 : M > 0,$$

$$H_0 : M = 0 \quad - \quad H_1 : M < 0.$$

Η μηδενική υπόθεση μπορεί να γραφεί και ως $H_0 : p = 0.5$, όπου $p = P(X > Y)$. Σημειώνουμε ότι η μηδενική υπόθεση δεν ισχυρίζεται ότι οι δύο πληθυσμοί X και Y έχουν την ίδια διάμεσο. Πράγματι δεν είναι πάντα αληθές ότι $M = M_X - M_Y$ (ισχύει όμως στην περίπτωση που οι δύο πληθυσμοί X, Y είναι συμμετρικοί). Η διάμεσος M αναφέρεται συνήθως ως «επίδραση της θεραπείας» (treatment effect) αφού το κλασικό πεδίο εφαρμογής των παραπάνω ελέγχων είναι η περίπτωση που τα X_i και τα Y_i είναι μετρήσεις μιας ποσότητας πριν και μετά τη χορήγηση κάποιας θεραπείας, αντίστοιχα. Η μηδενική υπόθεση ισχυρίζεται ότι δεν εμφανίζεται καμία "μετατόπιση" του αρχικού πληθυσμό λόγω της θεραπείας.

Η στατιστική συνάρτηση ελέγχου S ορίζεται ως ο συνολικός αριθμός των θετικών διαφορών $X_1 - Y_1$, $X_2 - Y_2$, ..., $X_n - Y_n$. Μεγάλες (αντίστοιχα μικρές) τιμές της S οδηγούν στο συμπέρασμα ότι θα πρέπει να ισχύει $M > 0$ (αντίστοιχα $M < 0$). Προφανώς η S ακολουθεί κατανομή $B(n, 0.5)$ όταν η μηδενική υπόθεση είναι αληθής. Αν υπάρχουν μηδενικές διαφορές τότε συνήθως τις παρα-

λείπουμε και μειώνεται ανάλογα το μέγεθος του δείγματος n. Αν συμβολίσουμε με s την παρατηρούμενη τιμή της *S*, τότε ο υπολογισμός της $p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	$p-value$
$H_1: M < 0$	$P(S \le s) = \sum_{i=0}^{s} \binom{n}{i} 2^{-n}$
$H_1: M > 0$	$p-value = P(S \ge s) = \sum_{i=s}^{n} \binom{n}{i} 2^{-n}$
$H_1: M \ne 0$	$\sum_{i=0}^{n} I[P(S=i) \le P(S=s)] \binom{n}{i} 2^{-n}$

Εναλλακτικά, οι παραπάνω έλεγχοι στην περίπτωση που το μέγεθος του δείγματος είναι σχετικά μεγάλο μπορούν να εκτελεστούν με την κανονική προσέγγιση της διωνυμικής κατανομής. Η στατιστική συνάρτηση ελέγχου τότε είναι η

$$Z = \frac{S - E(S)}{\sqrt{Var(S)}} = \frac{S - n/2}{\sqrt{n}/2}.$$

Αν συμβολίσουμε με z την παρατηρούμενη τιμή της *Z*, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a και η $p-value$ του ελέγχου δίνονται στο ακόλουθο πλαίσιο

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**	$p-value$
$H_1: M < 0$	$\{z : z < -z_a\}$	$P(Z \le z)$
$H_1: M > 0$	$\{z : z > z_a\}$	$P(Z \ge z)$
$H_1: M \ne 0$	$\{z : z < -z_{a/2}\} \cup \{z : z > z_{a/2}\}$	$2P(Z \ge \lvert z \rvert) = 2(1 - P(Z \le \lvert z \rvert))$

Ένα "μειονέκτημα" του προσημικού ελέγχου είναι ότι δεν λαμβάνει υπόψη το μέγεθος των διαφορών $D_i = X_i - Y_i$, παρά μόνο πόσες από αυτές είναι θετικές. Ο έλεγχος Wilcoxon signed-rank για δείγματα κατά ζεύγη λαμβάνει υπόψη το μέγεθος των διαφορών αντιστοιχώντας μεγάλο βαθμό (rank) στις μεγαλύτερες διαφορές, και μικρό στις μικρότερες. Ο έλεγχος προσήμου μπορεί να εφαρμοστεί γενικότερα σε διατάξιμες παρατηρήσεις (ordinal data). Επίσης ο έλεγχος προσήμου μπορεί να χρησιμοποιηθεί αντί του ελέγχου για τη διαφορά των μέσων τιμών με δείγματα κατά ζεύγη (paired t-test) όταν αμφιβάλλουμε για την κανονικότητα των διαφορών $D_i = X_i - Y_i$, $i = 1,2,...,n$.

Ο προσημικός έλεγχος γίνεται στο R με τη συνάρτηση `SIGN.test` του πακέτου BSDA (επίσης και με το πακέτο PASWR). Η βασική σύνταξη της συνάρτησης `SIGN.test` είναι η ακόλουθη

```
SIGN.test(x, y, md=M0)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
md: η τιμή της διαμέσου Μ των διαφορών υπό τη μηδενική υπόθεση
```

Παράδειγμα 6.27 (sign.test με δείγματα κατά ζεύγη)

Η μείωση της ροής του αίματος στον εγκέφαλο στους ηλικιωμένους ανθρώπους μπορεί να οδηγήσει σε χειροτέρευση της διανοητικής τους κατάστασης. Σε 11 ηλικιωμένα άτομα χορηγήθηκε η φαρμακευτική ουσία κυκλανδελάτη με σκοπό να αυξηθεί η ταχύτητα ροής του αίματος στον εγκέφαλο. Στον ακόλουθο πίνακα δίνεται ο χρόνος (σε sec) που χρειάζεται για να μεταφερθεί αίμα από την καρωτίδα αρτηρία στη σφαγίτιδα φλέβα πριν (X) και 4 μήνες μετά (Y) τη χορήγηση κυκλανδελάτης σε καθημερινή βάση στους 11 ηλικιωμένους.

X	15	12	12	14	13	13	13	12	12.5	12	12.5
Y	13	8	12.5	12	12	12.5	12.5	14	12	11	10

Αν όντως η κυκλανδελάτη αυξάνει τη ροή του αίματος προς τον εγκέφαλο τότε αυτό συνεπάγεται μειωμένο χρόνο μεταφοράς του αίματος από την καρωτίδα αρτηρία στη σφαγίτιδα φλέβα. Για να ελέγξουμε αν η κυκλανδελάτη αυξάνει τη ροή του αίματος θα πρέπει να εκτελέσουμε τον έλεγχο

$$H_0 : M = 0 \quad - \quad H_1 : M > 0 .$$

Χρησιμοποιώντας το R παίρνουμε

```
> library(BSDA)
> x <- c(15,12,12,14,13,13,13,12,12.5,12,12.5)
> y <- c(13,8,12.5,12,12,12.5,12.5,14,12,11,10)
> SIGN.test(x, y, alternative="greater")

        Dependent-samples Sign-Test

data:  x and y
S = 9, p-value = 0.03271
alternative hypothesis: true median difference is greater than 0
95 percent confidence interval:
 0.5 Inf
sample estimates:
median of x-y
            1

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.8867    0.5    Inf
Interpolated CI       0.9500    0.5    Inf
Upper Achieved CI     0.9673    0.5    Inf
```

Επιβεβαίωση της *p-value*:

```
> d <- x-y
> w <- ifelse(d>0, 1, 0)
> s <- sum(w); n <- length(x); p <- 0.5
> p.value <- sum(dbinom(s:n,n,p))
```

```
> cat("Επιβεβαίωση: p-value =", p.value, "\n")
Επιβεβαίωση: p-value = 0.03271484
```

Επομένως σε επίπεδο σημαντικότητας $a = 0.05$ η μηδενική υπόθεση απορρίπτεται, οπότε η κυκλανδελάτη μειώνει το χρόνο μεταφοράς του αίματος από την καρωτίδα αρτηρία στη σφαγίτιδα φλέβα. ■

6.3.9 Έλεγχοι Wilcoxon για δύο δείγματα

6.3.9.1 Έλεγχος Wilcoxon Signed-Rank με δείγματα κατά ζεύγη

Έστω ένα τυχαίο δείγμα ζευγαρωτών παρατηρήσεων (X_i, Y_i), $1 \le i \le n$. Υποθέτουμε ότι οι διαφορές $D_i = X_i - Y_i$ είναι ανεξάρτητες τυχαίες μεταβλητές προερχόμενες από ένα συνεχή πληθυσμό D ο οποίος είναι συμμετρικός γύρω από τη διάμεσό του M. Αν συμβολίσουμε με F τη συνάρτηση κατανομής του D, τότε

$$F(M+t)+F(M-t)=1, \quad \text{για κάθε } t \in R.$$

Μας ενδιαφέρουν οι έλεγχοι

$$H_0 : M = 0 \quad - \quad H_1 : M \neq 0,$$

$$H_0 : M = 0 \quad - \quad H_1 : M > 0,$$

$$H_0 : M = 0 \quad - \quad H_1 : M < 0.$$

Η διάμεσος M αναφέρεται συνήθως ως «επίδραση της θεραπείας» (treatment effect) αφού το κλασικό πεδίο εφαρμογής των παραπάνω ελέγχων αναφέρεται στην περίπτωση που τα X_i και τα Y_i είναι μετρήσεις μιας ποσότητας πριν και μετά τη χορήγηση κάποιας θεραπείας, αντίστοιχα. Η μηδενική υπόθεση ισχυρίζεται ότι η κατανομή των διαφορών $D_i = X_i - Y_i$ είναι συμμετρική γύρω από το 0 που συνεπάγεται ότι δεν εμφανίζεται καμία "μετατόπιση" του αρχικού πληθυσμό λόγω της θεραπείας. Η εναλλακτική υπόθεση $M > 0$ ($M < 0$) δηλώνει ότι ο πληθυσμός X έχει μεγαλύτερη (μικρότερη) διάμεσο από τον πληθυσμό Y.

Για να εκτελεστούν οι παραπάνω έλεγχοι εφαρμόζουμε τη μεθοδολογία της Παραγράφου 6.2.5. στις διαφορές $D_i = X_i - Y_i$, $1 \le i \le n$.

Στο R ο έλεγχος Wilcoxon Signed-Rank για ζευγαρωτές παρατηρήσεις πραγματοποιείται με τη συνάρτηση `wilcox.test`. κάνοντας χρήση του ορίσματος `paired`. Η βασική σύνταξη της συνάρτησης `wilcox.test` για ζευγαρωτές παρατηρήσεις είναι η ακόλουθη

```
wilcox.test(x, y, paired=TRUE, exact=TRUE or FALSE, correct=TRUE
    or FALSE)
x, y: τα δείγματα σε μορφή διανύσματος
exact: ακριβής (TRUE) ή προσεγγιστικός (FALSE) υπολογισμός της
    p.value (αν δεν δηλωθεί το όρισμα exact τότε για n<50 παίρνου-
    με ακριβή p-value)
correct: υπολογισμός p-value με διόρθωση (TRUE) ή χωρίς διόρθωση
    (FALSE) συνέχειας (αν δεν δηλωθεί το όρισμα correct έχει de-
    fault τιμή TRUE)
```

Παράδειγμα 6.28 (wilcox.test για ζευγαρωτές παρατηρήσεις – δίπλευρη εναλλακτική)

Ας θεωρήσουμε τα δύο ακόλουθα δείγματα

X_i	Y_i
3, 5, 5, 4, 3, 7, 8, 7	7, 6, 3, 8, 5, 9, 7, 9

για τα οποία μας ενδιαφέρει ο έλεγχος

$$H_0 : M = 0 \quad - \quad H_1 : M \neq 0 .$$

Θα εργαστούμε όπως στην Παράγραφο 6.2.5. Κατασκευάζουμε τον ακόλουθο πίνακα

i	X_i	Y_i	D_i	$\lvert D_i \rvert$	R_i	k_i	$k_i R_i$	$(1-k_i)R_i$
1	3	7	-4	4	7.5	0	0	7.5
2	5	6	-1	1	1.5	0	0	1.5
3	5	3	2	2	4.5	1	4.5	0
4	4	8	-4	4	7.5	0	0	7.5
5	3	5	-2	2	4.5	0	0	4.5
6	7	9	-2	2	4.5	0	0	4.5
7	8	7	1	1	1.5	1	1.5	0
8	7	9	-2	2	4.5	0	0	4.5
					36		$T^+ = 6$	$T^- = 30$

Αφού $n = 8$ και $t^+ = 6$ έχουμε

$$t^* = \frac{t^+ - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{3}(t_j^3 - t_j)}} = \frac{6-18}{\sqrt{51 - \frac{(2^3-2)+(2^3-2)+(4^3-4)}{48}}} = -1.70560753 .$$

Συνεπώς η προσεγγιστική τιμή της $p-value$ του ελέγχου είναι ίση με $2(1-\Phi(|t^*|) = 0.08808151$.

Χρησιμοποιώντας το R παίρνουμε

```
> x <- c(3, 5, 5, 4, 3, 7, 8, 7)
> y <- c(7, 6, 3, 8, 5, 9, 7, 9)
> wilcox.test(x, y, paired=TRUE, exact=FALSE, correct=FALSE)

        Wilcoxon signed rank test
```

```
data:  y and x
V = 6, p-value = 0.08808
alternative hypothesis: true location shift is not equal to 0
```

Επιβεβαίωση της *p-value*:

```
> cat("Επιβεβαίωση: p-value=", 2*(1-pnorm(1.70560573)), "\n")
Επιβεβαίωση: p-value= 0.08808151
```

Στο ίδιο φυσικά αποτέλεσμα καταλήγουμε και με τη χρήση της στατιστικής συνάρτηση ελέγχου T^-. Πράγματι

```
> wilcox.test(y, x, paired=TRUE, exact=FALSE, correct=FALSE)

        Wilcoxon signed rank test

data:  y and x
V = 30, p-value = 0.08808
alternative hypothesis: true location shift is not equal to 0
```

Παράδειγμα 6.29 (wilcox.test για ζευγαρωτές παρατηρήσεις – μονόπλευρη εναλλακτική)

Σε ένα πείραμα μετρήθηκαν τα επίπεδα της ανοσοαντιδραστικής ινσουλίνης στο αίμα 7 σκύλων διαφόρων βαρών πριν (X) και 5 λεπτά μετά (Y) από τεχνητό ερεθισμό των νεύρων στα σπλάχνα. Οι μετρήσεις δίνονται στον ακόλουθο πίνακα

X	350	200	240	290	90	370	240
Y	480	130	250	310	280	1450	280

Θέλουμε να διαπιστώσουμε αν ο ερεθισμός των νεύρων αυξάνει τα επίπεδα της ανοσοαντιδραστικής ινσουλίνης στο αίμα, δηλαδή μας ενδιαφέρει ο έλεγχο της υπόθεσης

$$H_0 : M = 0 \quad - \quad H_1 : M < 0 .$$

Χρησιμοποιώντας το R παίρνουμε

```
> x <- c(350, 200, 240, 290, 90, 370, 240)
> y <- c(480, 130, 250, 310, 280, 1450, 280)
> wilcox.test(x, y, paired=TRUE, alternative="less", exact=TRUE)

        Wilcoxon signed rank test

data:  x and y
V = 4, p-value = 0.05469
alternative hypothesis: true location shift is less than 0
```

Επομένως σε επίπεδο σημαντικότητας $a \geq 0.055$ απορρίπτουμε τη μηδενική υπόθεση και καταλήγουμε στο συμπέρασμα ότι ο ερεθισμός των νεύρων αυξάνει τα επίπεδα της ανοσοαντιδραστικής ινσουλίνης στο αίμα των σκύλων.

6.3.9.2 Έλεγχος Wilcoxon Rank-Sum ή έλεγχος Mann-Whitney U

Ο έλεγχος Wilcoxon Rank-Sum (ή έλεγχος Mann-Whitney U) για δύο ανεξάρτητα τυχαία δείγματα

χρησιμοποιείται, όπως και ο έλεγχος Kolmogorov-Smirnov, για να ελέγξουμε αν αυτά προέρχονται από την ίδια συνεχή κατανομή χωρίς αυτή να προσδιορίζεται. Έτσι αν $X_1, X_2, ..., X_n$ και $Y_1, Y_2, ..., Y_m$ είναι τα δύο δείγματα, μας ενδιαφέρουν οι έλεγχοι

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \le F_Y(x) \quad (\text{ή } H_0 : \theta = 0 \quad - \quad H_1 : \theta < 0)$$

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \ge F_Y(x) \quad (\text{ή } H_0 : \theta = 0 \quad - \quad H_1 : \theta > 0)$$

$$H_0 : F_X(x) = F_Y(x) \quad - \quad H_1 : F_X(x) \ne F_Y(x) \quad (\text{ή } H_0 : \theta = 0 \quad - \quad H_1 : \theta \ne 0)$$

(σε παρένθεση δίνεται ο έλεγχος για το location model που είναι το σύνηθες πεδίο εφαρμογής του συγκεκριμένου ελέγχου). Η στατιστική συνάρτηση ελέγχου W δίνεται από τη σχέση

$$W = \sum_{j=1}^{m} S_j$$

όπου S_j είναι ο βαθμός (rank) της παρατήρησης Y_j, $j = 1, 2, ..., m$, στη διάταξη των $n+m$ παρατηρήσεων των δύο δειγμάτων από τη μικρότερη παρατήρηση έως τη μεγαλύτερη. Είναι προφανές ότι οι βαθμοί που αντιστοιχούν στα Y_j θα είναι γενικά μεγαλύτεροι από αυτούς που αντιστοιχούν στα X_i, $i = 1, 2, ..., n$, όταν $M_Y > M_X$ ή όταν ισοδύναμα $\theta > 0$. Έτσι ο Wilcoxon πρότεινε να αποδεχόμαστε την εναλλακτική υπόθεση $H_1 : \theta > 0$ για μεγάλες τιμές του W. Ανάλογα θα πρέπει να αποδεχόμαστε την εναλλακτική υπόθεση $H_1 : \theta < 0$ για μικρές τιμές του W. Ο ακόλουθος πίνακας δίνει την κρίσιμη περιοχή των ελέγχων σε επίπεδο σημαντικότητας a

Εναλλακτική υπόθεση	**Κρίσιμη περιοχή**
$H_1 : \theta < 0$	$\{w : w \le w_a\}$
$H_1 : \theta > 0$	$\{w : w \ge w'_a\}$
$H_1 : \theta \ne 0$	$\{w : w \le w_{a/2}\} \bigcup \{w : w \ge w'_{a/2}\}$

Οι τιμές w_a και w'_a που εμφανίζονται στο παραπάνω πλαίσιο υπολογίζονται κάνοντας χρήση της ακριβής κατανομής του W σύμφωνα με τις σχέσεις $P(W \le w_a) \le a$ και $P(W \ge w'_a) \le a$.

Η κατανομή της W (διακριτή κατανομή) είναι γνωστή ως Wilcoxon Rank-Sum κατανομή και είναι συμμετρική με

$$E(W) = \frac{m(n+m+1)}{2}, \quad Var(W) = \frac{nm(n+m+1)}{12}.$$

Αποδεικνύεται ότι η στατιστική συνάρτηση

$$W^* = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{W - m(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}$$

για μεγάλα n, m (>10) ακολουθεί προσεγγιστικά την κατανομή $N(0,1)$ όταν η H_0 είναι αληθής. Αν συμβολίσουμε με $w*$ την παρατηρούμενη τιμή της $W*$, τότε ο προσεγγιστικός υπολογισμός της $p-value$ του ελέγχου γίνεται σύμφωνα με τον ακόλουθο πίνακα

Εναλλακτική υπόθεση	***p – value***
$H_1: \theta < 0$	$P(W* \le w*) = \Phi(w*)$
$H_1: \theta > 0$	$P(W* \ge w*) = 1 - \Phi(w*)$
$H_1: \theta \ne 0$	$2P(W* \ge \lvert w* \rvert) = 2(1 - \Phi(\lvert w* \rvert)$

Αν υπάρχουν δεσμοί (ισοπαλίες, ties) στους βαθμούς τότε χρησιμοποιείται η στατιστική συνάρτηση

$$W* = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{W - m(n+m+1)/2}{\sqrt{\frac{nm}{12}\left(n+m+1-\sum_{j=1}^{r}\frac{(t_j^3 - t_j)}{(n+m)(n+m-1)}\right)}}$$

η οποία ακολουθεί ασυμπτωτικά ($n, m \to \infty$) την κατανομή $N(0,1)$ όταν η H_0 είναι αληθής. Στον παραπάνω τύπο το r δηλώνει το πλήθος των διαφορετικών βαθμών, και το t_j δηλώνει πόσες φορές εμφανίζεται κάθε διαφορετικός βαθμός $(1 \le j \le r)$. Παρατηρούμε ότι η παρουσία των δεσμών δεν επηρεάζει τη μέση τιμή της W αλλά επηρεάζει τη διακύμανσή της.

Οι παραπάνω έλεγχοι μπορούν να εκτελεστούν και με την επονομαζόμενη Mann-Whitney U στατιστική συνάρτηση ελέγχου η οποία συνδέεται με τη στατιστική συνάρτηση W με τη σχέση (περίπτωση χωρίς δεσμούς)

$$U = W - \frac{m(m+1)}{2}.$$

Μπορεί να αποδειχθεί ότι $U = \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i < Y_j)$, δηλαδή ότι η τιμή της U είναι το πλήθος των X_i που είναι μικρότερα από το Y_1, συν το πλήθος των X_i που είναι μικρότερα από το Y_2, ..., συν το πλήθος των X_i που είναι μικρότερα από το Y_m. Κάτω από την H_0 η στατιστική συνάρτηση

$$U* = \frac{U - E(U)}{\sqrt{Var(U)}} = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}} \quad (= W*)$$

ακολουθεί ασυμπτωτικά ($n, m \to \infty$) την κατανομή $N(0,1)$.

Ολοκληρώνοντας τη συζήτηση του ελέγχου Wilcoxon Rank-Sum σημειώνουμε ότι για την εφαρμογή του απαιτείται να έχουμε μόνο συνεχείς κατανομές και ανεξάρτητα δείγματα. Ωστόσο εφαρμόζεται κάλλιστα και για διατάξιμα δεδομένα (ordinal data). Ο έλεγχος Wilcoxon Rank-Sum μπορεί να χρησιμοποιηθεί για το έλεγχο των μέσων ή των διαμέσων δύο πληθυσμών σε περιπτώσεις που είναι λογικό να χρησιμοποιηθεί το location model για την περιγραφή των υποθέσεών μας (μη-

δενικής και εναλλακτικής) και σε αυτές τις περιπτώσεις θεωρείται το καλύτερο μη παραμετρικό τεστ. Αυτός είναι άλλωστε ο λόγος που χρησιμοποιείται αντί του t-test για τη σύγκριση των μέσων δύο ανεξάρτητων δειγμάτων όταν αμφιβάλλουμε για την κανονικότητας των παρατηρήσεων.

Στο R ο έλεγχος Wilcoxon Rank-Sum για δύο δείγματα πραγματοποιείται με τη συνάρτηση `wilcox.test`. Η βασική σύνταξη της συνάρτησης `wilcox.test` για δύο δείγματα είναι η ακόλουθη

```
wilcox.test(y, x)
x: το πρώτο δείγμα σε μορφή διανύσματος
y: το δεύτερο δείγμα σε μορφή διανύσματος
```

Προσέξτε ότι πρώτα δηλώνουμε το δεύτερο δείγμα και μετά το πρώτο στη συνάρτηση `wilcox.test` για δύο δείγματα.

Παράδειγμα 6.29 (wilcox.test – μονόπλευρη εναλλακτική)

Ένας μηχανικός σε μια μεταλλουργία υποπτεύεται ότι ο χρόνος που χρειάζεται ένας εργάτης για να γεμίσει με λιωμένο μέταλλο ένα συγκεκριμένο καλούπι εξαρτάται από το αν η συγκεκριμένη εργασία εκτελεστεί πριν (X) ή μετά (Y) από το διάλειμμα για κολατσιό. Ο μηχανικός συνέλεξε 11 παρατηρήσεις που δίνονται στο ακόλουθο πλαίσιο

X						Y				
12.6	11.2	11.4	9.4	13.2	12.0	15.4	14.1	14.0	13.4	11.3

Ο μηχανικός υποπτεύεται ότι οι χρόνοι χύτευσης πριν το γεύμα είναι μικρότεροι από τους χρόνους χύτευσης μετά το γεύμα. Ο έλεγχος που μας ενδιαφέρει είναι ο

$$H_0 : \theta = 0 \quad - \quad H_1 : \theta > 0 .$$

Έχουμε ότι

Πληθυσμός	X	X	Y	X	X	X	X	Y	Y	Y	Y
Διάταξη	9.4	11.2	11.3	11.4	12.0	12.6	13.2	13.4	14.0	14.1	15.4
Βαθμός	1	2	**3**	4	5	6	7	**8**	**9**	**10**	**11**

Από τα παραπάνω παίρνουμε ότι

$$n = 6 , \quad m = 5 , \quad w = 3 + 4 + 8 + 9 + 10 + 11 = 41 , \quad u = 41 - 5 \cdot 6 / 2 = 26 .$$

Επίσης

$$w^* = u^* = \frac{w - m(n+m+1)/2}{\sqrt{nm(n+m+1)/12}} = \frac{26 - 5 \cdot (6+5+1)/2}{\sqrt{6 \cdot 5 \cdot (6+5+1)/12}} = 2.00831604 .$$

Συνεπώς $p - value = 1 - \Phi(w^*) = 0.02230486$.

Με το R παίρνουμε

```
> x <- c(12.6, 11.2, 11.4, 9.4, 13.2, 12.0)
> y <- c(15.4, 14.1, 14.0, 13.4, 11.3)
```

```
> wilcox.test(y, x, alternative="greater")

        Wilcoxon rank sum test

data:  y and x
W = 26, p-value = 0.02597
alternative hypothesis: true location shift is greater than 0
```

Η τιμή της στατιστικής συνάρτησης ελέγχου είναι ίση με 26 (Mann-Whitney U) και η τιμή της $p-value$ είναι ακριβής (αφού υπάρχουν λιγότερα από 50 συνολικά δεδομένα χωρίς δεσμούς) με διόρθωση συνέχειας (το όρισμα `correct` παίρνει την προεπιλεγμένη τιμή `TRUE`).

Η $p-value$ του ελέγχου με κανονική προσέγγιση (`exact=FALSE`) χωρίς διόρθωση συνέχειας (`correct=FALSE`) προκύπτει ως ακολούθως

```
> wilcox.test(y, x, alternative="greater", exact=FALSE, correct=FALSE)

        Wilcoxon rank sum test

data:  y and x
W = 26, p-value = 0.02230
alternative hypothesis: true location shift is greater than 0
```

Επιβεβαίωση της *p-value*:

```
> u <-wilcox.test(y,x)$statistic
> n <- length(x)
> m <- length(y)
> ustar <- (u-n*m/2)/sqrt(n*m*(n+m+1)/12)
> cat("Επιβεβαίωση: p-value =", 1-pnorm(ustar), "\n")
Επιβεβαίωση: p-value = 0.02230486
```

Ο αντίστοιχος έλεγχος Kolmogorov-Smirnov δίνει τα ακόλουθα αποτελέσματα

```
> ks.test(x, y, alternative="greater")

        Two-sample Kolmogorov-Smirnov test

data:  x and y
D^+ = 0.8, p-value = 0.03047
alternative hypothesis: the CDF of x lies above that of y
```

Τελικά σε επίπεδο σημαντικότητας $a = 0.05$, ο μηχανικός είχε δίκιο. ■

6.4 Έλεγχοι χ^2

6.4.1 Έλεγχος χ^2 καλής προσαρμογής για διακριτές κατανομές

Θεωρούμε ένα πείραμα τύχης με k δυνατά αποτελέσματα E_i, $1 \le i \le k$. Μας ενδιαφέρει ο έλεγχος

$$H_0 : P(E_1) = p_1,\ P(E_2) = p_2,\ \ldots,\ P(E_k) = p_k \quad - \quad H_1 : \text{Όχι } H_0$$

($p_1 + p_2 + \ldots + p_k = 1$). Σε n ανεξάρτητες επαναλήψεις του πειράματος θεωρούμε την τυχαία μεταβλητή n_i που δηλώνει πόσες φορές εμφανίστηκε φορές το ενδεχόμενο E_i, $1 \le i \le k$ ($n_1 + n_2 + \ldots + n_k = n$). Η στατιστική συνάρτηση ελέγχου είναι η

$$U = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

η οποία ακολουθεί προσεγγιστικά την κατανομή χ^2_{k-1} όταν η H_0 είναι αληθής (η προσέγγιση είναι καλή όταν $np_i \ge 5$ για $1 \le i \le k$). Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$$\{u : u \ge \chi^2_{k-1;\,a}\}$$

όπου u είναι η παρατηρούμενη τιμή της U. Η $p-value$ του ελέγχου είναι ίση με $P(\chi^2_{k-1} \ge u)$. Στην περίπτωση που κάποια από τα p_i είναι άγνωστα (έστω ότι s το πλήθος από τα p_i είναι άγνωστα) τότε θα πρέπει να αντικατασταθούν στην U με εκτιμητές μεγίστης πιθανοφάνειας (δηλαδή αντικαθιστώνται με τα $\hat{p}_i = n_i / n$). Σε αυτή την περίπτωση η U ακολουθεί προσεγγιστικά κατανομή χ^2_{k-s-1}.

Για τον παραπάνω έλεγχο στο R χρησιμοποιείται η συνάρτηση `chisq.test`. Η βασική σύνταξη της συνάρτησης `chisq.test` για τον προαναφερθέντα έλεγχο είναι η ακόλουθη

```
chisq.test(x, p=c(p1,...,pk), simulate.p.value =TRUE or FALSE,
    B=2000)
x: το διάνυσμα (n1,...,nk)
p: το διάνυσμα (p1,...,pk)
simulate.p.value: Το όρισμα TRUE δηλώνει αν θα χρησιμοποιηθεί προ-
    σομοίωση Monte Carlo για τον υπολογισμό της p.value (default
    τιμή FALSE)
B: δηλώνει τον αριθμό των επαναλήψεων της προσομοίωσης
```

Παράδειγμα 6.30 (chisq.test – πολυωνυμική κατανομή)

Για παράδειγμα, έστω ότι ένα πείραμα τύχης εκτελείται 186 φορές και τα 5 δυνατά αποτελέσματά του E_i, $1 \le i \le 5$, εμφανίζονται 89, 37, 30, 28 και 2 φορές αντίστοιχα. Για τον έλεγχο της υπόθεσης

$$H_0 : p_1 = 0.4,\ p_2 = 0.2,\ p_3 = 0.2,\ p_4 = 0.18,\ p_5 = 0.02 \quad - \quad H_1 : \text{Όχι } H_0$$

με το R, όπου $p_i = P(E_i)$, έχουμε τα ακόλουθα

```
> x <- c(89,37,30,28,2)
> px <- c(0.40,0.20,0.20,0.18,0.02)
> chisq.test(x, p = px)

        Chi-squared test for given probabilities
```

```
data:  x
X-squared = 5.9519, df = 4, p-value = 0.2028

Warning message:
In chisq.test(x, p = p) : Chi-squared approximation may be incorrect
```

Επιβεβαίωση της *p-value*:

```
> u <- sum((x-186*px)^2/(186*px))
> cat("Επιβεβαίωση: p-value =", 1-pchisq(u, df=length(x)-1), "\n")
Επιβεβαίωση: p-value = 0.2027684
```

Παρατηρούμε ότι υπάρχει προειδοποιητικό μήνυμα ότι η χ^2 προσέγγιση μπορεί να μην είναι αξιόπιστη αφού παραβιάζεται η συνθήκη $np_i \geq 5$ για τουλάχιστον ένα i $(1 \leq i \leq 5)$. Σε αυτή την περίπτωση μια εναλλακτική λύση, πέραν της κλασικής λύσης της ομαδοποίησης των δεδομένων (δείτε το επόμενο παράδειγμα), είναι να χρησιμοποιηθεί ένα Monte Carlo τεστ που πρότεινε ο Hope (1968)[‡] με χρήση του ορίσματος `simulate.p.value`. Έτσι

```
> sum(x)*px
[1] 74.40 37.20 37.20 33.48  3.72
# np[5]=3.72<5
> chisq.test(x, p = px, simulate.p.value = TRUE, B=2000)

        Chi-squared test for given probabilities with simulated p-value
(based on 2000 replicates)

data:  x
X-squared = 5.9519, df = NA, p-value = 0.2014
```

Ο έλεγχος που περιγράψαμε παραπάνω δύναται να χρησιμοποιηθεί για να ελέγξουμε αν κάποια δεδομένα προέρχονται από κάποια συγκεκριμένη διακριτή, ή ακόμη και συνεχή κατανομή, δηλαδή για τον έλεγχο

$$H_0 : F(x) = F_0(x) \quad - \quad H_1 : F(x) \neq F_0(x).$$

Ωστόσο, στην συνεχή περίπτωση, προτιμώνται συνήθως άλλες πιο κομψές μέθοδοι (π.χ. ο έλεγχος Kolmogorov-Smirnov της Παραγράφου 6.2.7.1). Για τη διακριτή περίπτωση δίνουμε το ακόλουθο παράδειγμα.

Παράδειγμα 6.31 (chisq.test – κατανομή Poisson)

Στον ακόλουθο πίνακα δίνεται ο αριθμός των γκολ που μπήκαν στους τελευταίους $n = 232$ ποδοσφαιρικούς αγώνες της SuperLeague

# Γκολ (i)	0	1	2	3	4	5	6	7	8
# Αγώνων (n_i)	19	49	60	47	32	18	3	3	1

[‡] Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure, *J. Roy, Statist. Soc. B* **30**, 582–598.

Θέλουμε να ελέγξουμε αν τα δεδομένα προέρχονται από την κατανομή Poisson με $\lambda = 2.5$.

Εργαζόμαστε με το R ως εξής: Αρχικά οι τιμές n_i, $i = 0, 1, ..., 8$, αποθηκεύονται στο διάνυσμα `Obs1`. Στη συνέχεια υπολογίζουμε τις πιθανότητες $p_i = P(X = i)$, $0 \le i \le 7$, και την πιθανότητα $p_8 = P(X \ge 8)$ για $X \sim P(2.5)$ (δείτε το διάνυσμα `Prob1`). Σημειώνουμε ότι $p_0 + p_1 + ... + p_8 = 1$. Στη συνέχεια υπολογίζουμε τις ποσότητες np_i (δείτε το διάνυσμα `Exp1`). Ο έλεγχος που μας ενδιαφέρει είναι ο

$$H_0 : P(E_0) = p_0,\ P(E_1) = p_1,\ \ldots,\ P(E_8) = p_8 \quad - \quad H_1 : \text{Όχι } H_0$$

όπου E_i, $0 \le i \le 7$, είναι το ενδεχόμενο να σημειωθούν i γκολ σε ένα ματς και E_8 είναι το ενδεχόμενο να σημειωθούν τουλάχιστον 8 γκολ σε ένα ματς. Μη απόρριψη της H_0 οδηγεί στο συμπέρασμα ότι τα δεδομένα προέρχονται από την κατανομή $P(2.5)$.

```
> a <- c(0:8)
> Obs1 <- c(19,49,60,47,32,18,3,3,1)
> goals <- rep(a, Obs1)
> table(goals)
goals
 0  1  2  3  4  5  6  7  8
19 49 60 47 32 18  3  3  1
> Prob1 <- c(dpois(0:7,2.5),1-ppois(7,2.5)); Exp1 <- 232*Prob1
> ans1 <- cbind(Prob1,Exp1,Obs1)
> row.names(ans1) <- c(" X=0", " X=1", " X=2", " X=3", " X=4", " X=5",
+ " X=6"," X=7"," X>=8")
> ans1
            Prob1       Exp1 Obs1
 X=0   0.082084999 19.0437197   19
 X=1   0.205212497 47.6092992   49
 X=2   0.256515621 59.5116240   60
 X=3   0.213763017 49.5930200   47
 X=4   0.133601886 30.9956375   32
 X=5   0.066800943 15.4978188   18
 X=6   0.027833726  6.4574245    3
 X=7   0.009940617  2.3062230    3
 X>=8  0.004246695  0.9852334    1
```

Παρατηρούμε ότι $np_7 < 5$ και $np_8 < 5$ οπότε ο έλεγχος χ^2 καλής προσαρμογής δεν θα είναι αξιόπιστος. Αποφασίζουμε να ενοποιήσουμε τις παρατηρήσεις 6, 7, 8 και να κάνουμε τον έλεγχο

$$H_0 : P(E_0) = p_0,\ P(E_1) = p_1,\ \ldots,\ P(E_6) = p_6 \quad - \quad H_1 : \text{Όχι } H_0$$

όπου $p_i = P(X = i)$, $0 \le i \le 5$, και $p_6 = P(X \ge 6)$ για $X \sim P(2.5)$. Δουλεύοντας όπως πριν ορίζουμε το διάνυσμα `Prob` και υπολογίζουμε τα διανύσματα `Exp` και `Obs`.

```
> Prob <- c(dpois(0:5,2.5),1-ppois(5,2.5))
> Exp <- 232*Prob
> Obs <- c(19,49,60,47,32,18,7)
> ans <- cbind(Prob,Exp,Obs)
> row.names(ans) <- c(" X=0", " X=1", " X=2", " X=3", " X=4", " X=5",
```

```
+ "X>=6")
> ans
            Prob       Exp Obs
 X=0 0.08208500 19.04372  19
 X=1 0.20521250 47.60930  49
 X=2 0.25651562 59.51162  60
 X=3 0.21376302 49.59302  47
 X=4 0.13360189 30.99564  32
 X=5 0.06680094 15.49782  18
X>=6 0.04202104  9.74888   7
```

Παρατηρούμε ότι $np_i > 5$ για $0 \le i \le 6$, οπότε προχωρούμε στην εκτέλεση του ελέγχου.

```
> chisq.test(x=Obs,p=Prob)

        Chi-squared test for given probabilities

data:  Obs
X-squared = 1.3919, df = 6, p-value = 0.9663
```

Επιβεβαίωση της *p-value*:

```
> chi.obs <- sum((Obs-Exp)^2/Exp)
> df <- length(Obs)-1
> cat("Επιβεβαίωση: p-value=", 1-pchisq(chi.obs,df), "\n")
Επιβεβαίωση: p-value= 0.9663469
```

Επομένως καταλήγουμε στο συμπέρασμα ότι όντας τα δεδομένα μας προέρχονται από την κατανομή Poisson με $\lambda = 2.5$. Σημειώνεται ότι αν δεν δινόταν η τιμή του λ θα έπρεπε να εκτιμηθεί από τα δεδομένα και θα χρησιμοποιούσαμε για τον έλεγχο την κατανομή χ^2_{k-2}. ■

6.4.2 Έλεγχοι ανεξαρτησίας και ομογένειας σε πίνακες συνάφειας

Για να διερευνηθεί η ανεξαρτησία δύο κατηγορικών τυχαίων μεταβλητών X και Y με r $(A_1, A_2, ..., A_r)$ και k $(B_1, B_2, ..., B_k)$ δυνατές τιμές (επίπεδα, χαρακτηριστικά), αντίστοιχα, χρησιμοποιείται ένα χι-τετράγωνο τεστ παρόμοιο με αυτό της προηγούμενης παραγράφου. Ο έλεγχος που μας ενδιαφέρει είναι ο ακόλουθος

$$H_0 : p_{ij} = p_{i\bullet} p_{\bullet j} \quad 1 \le i \le r,\ 1 \le j \le k \quad - \quad H_1 : \text{Όχι } H_0$$

όπου

$p_{ij} = P(A_i B_j)$ είναι η πιθανότητα εμφάνισης του χαρακτηριστικού (A_i, B_j)

$p_{i\bullet} = P(A_i) = \sum_{j=1}^{k} p_{ij}$ είναι η πιθανότητα εμφάνισης του χαρακτηριστικού A_i

$p_{\bullet j} = P(B_j) = \sum_{i=1}^{r} p_{ij}$ είναι η πιθανότητα εμφάνισης του χαρακτηριστικού B_j

(προφανώς $\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} = \sum_{i=1}^{r} p_{i\bullet} = \sum_{j=1}^{k} p_{\bullet j} = 1$).

Από τον πληθυσμό παίρνουμε n παρατηρήσεις (X_i, Y_i), $1 \le i \le n$, και θεωρούμε τις τυχαίες μεταβλητές n_{ij} $(1 \le i \le r,\ 1 \le j \le k)$ που δηλώνουν πόσες φορές εμφανίστηκε το χαρακτηριστικό (A_i, B_j) στο δείγμα ($\sum_{i=1}^{r}\sum_{j=1}^{k} n_{ij} = n$). Ενδεικτικός είναι ο ακόλουθος πίνακας

Δειγματοληψία από ένα πληθυσμό – Δύο κριτήρια

	B_1	B_2	...	B_k	Σύνολο
A_1	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
A_2	n_{21}	n_{22}		n_{2k}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\bullet}$
Σύνολο	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	n

Προφανώς οι rk τυχαίες μεταβλητές n_{ij} $(1 \le i \le r,\ 1 \le j \le k)$ έχουν από κοινού κατανομή την πολυωνυμική με παραμέτρους n και p_{ij}. Σημειώνουμε επίσης ότι $E(n_{ij}) = np_{ij}$. Για τον έλεγχο

$$H_0 : p_{ij} = p_{i\bullet}p_{\bullet j} \quad 1 \le i \le r,\ 1 \le j \le k \quad - \quad H_1 : \text{Όχι } H_0$$

χρησιμοποιείται η στατιστική συνάρτηση ελέγχου

$$U = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(n_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

η οποία ακολουθεί προσεγγιστικά την κατανομή χ^2_{rk-1}, όταν η H_0 είναι αληθής (η προσέγγιση θεωρείται καλή αν όλα τα γινόμενα np_{ij} είναι μεγαλύτερα ή ίσα του 5). Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$$\{u : u \ge \chi^2_{rk-1;\,a}\}$$

όπου u είναι η παρατηρούμενη τιμή της U. Η $p-value$ του ελέγχου είναι ίση με $P(\chi^2_{rk-1} \ge u)$.

Στη συνήθη περίπτωση τα $p_{i\bullet}$ και $p_{\bullet j}$ είναι άγνωστα και αντικαθιστώνται στην U με τις ποσότητες (ΕΜΠ)

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} = \frac{\sum_{j=1}^{k} n_{ij}}{n}, \quad \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n} = \frac{\sum_{i=1}^{r} n_{ij}}{n}, \quad 1 \le i \le r,\ 1 \le j \ge k\,.$$

Σε αυτή την περίπτωση η U δίνεται από τον τύπο

$$U = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \dfrac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n}}$$

η οποία ακολουθεί προσεγγιστικά την κατανομή $\chi^2_{rk-1-(r-1)-(k-1)}$, δηλαδή την $\chi^2_{(r-1)(k-1)}$.

Για τον παραπάνω έλεγχο στο R χρησιμοποιείται η συνάρτηση `chisq.test`. Η βασική σύνταξη της συνάρτησης `chisq.test` για έλεγχο ανεξαρτησίας σε πίνακες συνάφειας είναι η ακόλουθη

```
chisq.test(m, correct=TRUE or FALSE, simulate.p.value = TRUE or
    FALSE, B = 2000)
m: πίνακας διαστάσεων rxk
correct: σε πίνακες 2x2 δηλώνει αν θα χρησιμοποιηθεί η διόρθωση
    συνέχειας του Yates με default τιμή TRUE (λειτουργεί μόνο όταν
    simulate.p.value=FALSE)
simulate.p.value: Το όρισμα TRUE δηλώνει αν θα χρησιμοποιηθεί προ-
    σομοίωση Monte Carlo για τον υπολογισμό της p.value (default
    τιμή FALSE)
B: δηλώνει τον αριθμό των επαναλήψεων της προσομοίωσης
```

Παράδειγμα 6.32 (chisq.test – έλεγχος ανεξαρτησίας)

Σε ένα τυχαίο δείγμα 100 ατόμων προέκυψαν τα ακόλουθα αποτελέσματα ως προς τις ιδιότητες φύλο και κάπνισμα

	Non-Smoker	Light Smoker	Heavy Smoker
Female	28	8	22
Male	26	2	14

Ο έλεγχος ανεξαρτησίας των χαρακτηριστικών φύλο και κάπνισμα του πληθυσμού γίνεται ως εξής:

```
> sex.smoke <- rbind(c(28, 8, 22), c(26, 2, 14))
> rownames(sex.smoke) <- c("Male", "Female")
> colnames(sex.smoke) <- c("Non-Smoker", "Light Smoker", "Heavy Smoker")
> sex.smoke
       Non-Smoker Light Smoker Heavy Smoker
Male           28            8           22
Female         26            2           14
> chisq.test(sex.smoke)

        Pearson's Chi-squared test

data:  sex.smoke
X-squared = 2.9678, df = 2, p-value = 0.2267

Warning message:
In chisq.test(sex.smoke) : Chi-squared approximation may be incorrect
```

Περισσότερες πληροφορίες για τον παραπάνω έλεγχο (όπως και για όλους τους ελέγχους) λαμβάνονται ως εξής:

```
> a <- chisq.test(sex.smoke)
> names(a)
[1] "statistic" "parameter" "p.value"   "method"    "data.name"
[6] "observed"  "expected"  "residuals"
> attach(a)
```

```
> observed
       Non-Smoker Light Smoker Heavy Smoker
Male            28             8           22
Female          26             2           14
> expected
       Non-Smoker Light Smoker Heavy Smoker
Male         31.32          5.8         20.88
Female       22.68          4.2         15.12
> residuals ## (observed - expected) / sqrt(expected)
       Non-Smoker Light Smoker Heavy Smoker
Male   -0.5932356    0.9135003    0.2451053
Female  0.6971345   -1.0734901   -0.2880329
```

Διαπιστώνουμε από τα παραπάνω ότι ένα γινόμενο $n_{i\bullet}n_{\bullet j}/n$ είναι μικρότερο του 5 (για την ακρίβεια 4.2), οπότε η προσέγγιση της σταιστικής συνάρτησης ελέγχου από την αντίστοιχη κατανομή χ^2 δεν είναι αρκετά καλή (υπάρχει σχετικό προειδοποιητικό μήνυμα).

Ορισμένες φορές υπάρχουν r διαφορετικοί πληθυσμοί των οποίων τα στοιχεία ταξινομούνται σε k διαφορετικές κατηγορίες ενός κριτηρίου. Ενδεικτικός είναι ο ακόλουθος πίνακας

Δειγματοληψία από *r* πληθυσμούς – Ένα κριτήριο

	Κατηγορία 1	Κατηγορία 2	…	Κατηγορία k	Σύνολο
Πληθυσμός 1	n_{11}	n_{12}	…	n_{1k}	$n_{1\bullet}$
Πληθυσμός 2	n_{21}	n_{22}	…	n_{2k}	$n_{2\bullet}$
⋮	⋮	⋮		⋮	⋮
Πληθυσμός r	n_{r1}	n_{r2}	…	n_{rk}	$n_{r\bullet}$
Σύνολο	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet k}$	$n_{\bullet\bullet}$

Τώρα, υποθέτουμε ότι τα αθροίσματα των γραμμών του παραπάνω πίνακα είναι σταθερά σε αντίθεση με τον πίνακα που χρησιμοποιήθηκε για τον έλεγχο ανεξαρτησίας, όπου το μόνο σταθερό ήταν το άθροισμα n όλων των στοιχείων του πίνακα.

Για παράδειγμα, σε ασθενείς που πάσχουν από μια συγκεκριμένη ασθένεια χορηγούνται δύο διαφορετικές θεραπευτικές αγωγές (οι δύο πληθυσμοί, $r=2$) και ταξινομούνται σε τρεις κατηγορίες ($k=3$) ανάλογα με το αν καλυτέρεψε, έμεινε στάσιμη ή αν χειροτέρεψε η κατάσταση της υγείας τους. Ενδιαφέρον θα είχε να μπορούσαμε να απαντήσουμε στην ερώτηση αν τα ποσοστά των ασθενών στις τρεις κατηγορίες είναι τα ίδια στους δύο πληθυσμούς. Στη γενική περίπτωση, ο έλεγχος που μας ενδιαφέρει είναι ο

$$H_0: p_{1j} = p_{2j} = \ldots = p_{rj} = p_j, \quad j = 1,2,\ldots,k \quad - \quad H_1: \text{Όχι } H_0$$

όπου p_{ij} είναι η πιθανότητα μια τυχούσα παρατήρηση από τον i πληθυσμό ($1 \le i \le r$) να κατανέμεται στην κατηγορία j ($1 \le j \le k$), και ονομάζεται έλεγχος ομογένειας. Τώρα οι k τυχαίες μετα-

βλητές $n_{i1}, n_{i2}, ..., n_{ik}$ $(1 \le i \le r)$ έχουν από κοινού κατανομή την πολυωνυμική με παραμέτρους $n_{i\bullet}$ και $p_{i1}, p_{i2}, ..., p_{ik}$. Η στατιστική συνάρτηση του ελέγχου δίνεται από τον τύπο

$$U = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(n_{ij} - n_{i\bullet}p_j)^2}{n_{i\bullet}p_j} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

η οποία ακολουθεί προσεγγιστικά την κατανομή $\chi^2_{r(k-1)}$, όταν η H_0 είναι αληθής. Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$$\{u : u \ge \chi^2_{r(k-1);a}\}$$

όπου u είναι η παρατηρούμενη τιμή της U. Η $p-value$ του ελέγχου είναι ίση με $P(\chi^2_{r(k-1)} \ge u)$.

Στη συνήθη περίπτωση τα p_j $(1 \le j \le k)$ είναι άγνωστα και εκτιμώνται με τις ποσότητες (ΕΜΠ)

$$\hat{p}_j = \frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{\sum_{i=1}^{r} n_{ij}}{n_{\bullet\bullet}}, \quad 1 \le j \le k.$$

Σε αυτή την περίπτωση η U δίνεται από τον τύπο

$$U = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}}$$

και ακολουθεί προσεγγιστικά κατανομή $\chi^2_{(r-1)(k-1)}$.

Για τον παραπάνω έλεγχο στο R χρησιμοποιείται και πάλι η συνάρτηση `chisq.test`.

Παράδειγμα 6.33 (chisq.test – έλεγχος ομογένειας)

Από το σύνολο των αποφοίτων μιας χρονιάς σε ένα κεντρικό πανεπιστήμιο επιλέχτηκαν τυχαία 100 φοιτητές και 200 φοιτήτριες. Η κατανομή του βαθμού του πτυχίου των 300 φοιτητών στις κατηγορίες "Καλώς", "Λίαν Καλώς" και "Άριστα" δίνεται στο ακόλουθο πίνακα

	Καλώς	Λίαν καλώς	Άριστα	Σύνολο
Φοιτητές	50	30	20	100
Φοιτήτριες	50	80	70	200
Σύνολο	100	110	90	300

Θέλουμε να ελέγξουμε αν η κατανομή του βαθμού του πτυχίου στους φοιτητές είναι διαφορετική από αυτή των φοιτητριών. Εκτελώντας τον έλεγχο με το R παίρνουμε τα ακόλουθα

```
> m <- matrix(c(50,50,30,80,20,70), nrow=2)
> chisq.test(m)
        Pearson's Chi-squared test
data:  m
X-squared = 19.3182, df = 2, p-value = 6.384e-05
```

Επιβεβαίωση της *p-value*:

```
> dimnames(m) <- list(SEX=c("Male", "Female"), GRADE=c("GOOD","VERY
GOOG", "EXCELLENT"))
> m
        GRADE
SEX      GOOD VERY GOOG EXCELLENT
  Male     50         30        20
  Female   50         80        70
> E <- outer(rowSums(m), colSums(m), "*")/sum(m)
> E
           GOOD VERY GOOG EXCELLENT
Male   33.33333  36.66667        30
Female 66.66667  73.33333        60
> chi.obs <- sum((m-E)^2/E)
> cat("Επιβεβαίωση: p-value =", 1-pchisq(chi.obs,2), "\n")
Επιβεβαίωση: p-value = 6.384253e-05
```

Επομένως το συμπέρασμα είναι ότι η κατανομή του βαθμού πτυχίου στις κατηγορίες "Καλώς", "Λίαν Καλώς" και "Άριστα" στους φοιτητές και στις φοιτήτριεσδεν είναι ίδια. ■

Παράδειγμα 6.34 (chisq.test και prop.test)

Στο Παράδειγμα 6.23 της Παραγράφου 6.3.5.2 χρησιμοποιήσαμε τα δεδομένα του ακόλουθου πίνακα

	Υπέρ	Κατά	Σύνολο
Άνδρες	100	900	1000
Γυναίκες	120	880	1000
Σύνολο	220	1780	2000

για να διαπιστώσουμε αν το ποσοστό p_1 των ανδρών και το ποσοστό p_2 των γυναικών που υποστηρίζουν τον πολιτικό διαφέρουν ή όχι. Εκτελέστηκε ο προσεγγιστικός έλεγχος της υπόθεσης

$$H_0 : p_1 - p_2 = 0 \quad - \quad H_1 : p_1 - p_2 \neq 0$$

με τη βοήθεια της συνάρτησης `prop.test`. Επαναλαμβάνουμε την ανάλυση αλλάζοντας την τιμή του ορίσματος `correct` από `FALSE` σε `TRUE`.

```
> prop.test(c(100,120), c(1000,1000), alternative="two.sided",
+ correct=TRUE)

    2-sample test for equality of proportions with continuity correction

data:  c(100, 120) out of c(1000, 1000)
X-squared = 1.8437, df = 1, p-value = 0.1745
alternative hypothesis: two.sided
```

Στην Παράγραφο 6.3.5.2 αναφέραμε ότι ο παραπάνω έλεγχος είναι ισοδύναμος με τον αντίστοιχο έλεγχο ομογένειας (ανεξαρτήτως της τιμής του ορίσματος `correct`). Πράγματι με τη συνάρτηση `chisq.test` προκύπτει το ίδιο αποτέλεσμα.

```
> m <-matrix(c(100,120,900,880), nrow=2)
> chisq.test(m, correct=TRUE)

        Pearson's Chi-squared test with Yates' continuity correction

data:  m
X-squared = 1.8437, df = 1, p-value = 0.1745
```

Για τον υπολογισμό της στατιστικής συνάρτησης ελέγχου U έχει χρησιμοποιηθεί η διόρθωση συνέχειας του Yates, δηλαδή

$$U = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}\right| - 0.5\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}}.$$

```
> E <- outer(rowSums(m), colSums(m), "*")/sum(m)
> chi.obs <- sum((abs(m-E)-0.5)^2/E); chi.obs
[1] 1.843718
> cat("Επιβεβαίωση: p-value=", 1-pchisq(chi.obs,1), "\n")
Επιβεβαίωση: p-value= 0.1745158
```

6.4.3 Έλεγχος ισότητας r αναλογιών

Ένα έλεγχος που η λογική του ακολουθεί τη λογική των δύο ελέγχων που παρουσιάσαμε στην προηγούμενη παράγραφο είναι ο έλεγχος ισότητας των αναλογιών r πληθυσμών, δηλαδή ο έλεγχος

$$H_0 : p_1 = p_2 = \ldots = p_r \quad - \quad H_1 : \text{Όχι } H_0$$

Το πρόβλημα μπορεί να παρουσιαστεί σε μορφή πινάκων συνάφειας $r \times 2$ στη μορφή

	Επιτυχίες	Αποτυχίες	Σύνολο
Πληθυσμός 1	x_1	$n_1 - x_1$	n_1
Πληθυσμός 2	x_2	$n_2 - x_2$	n_2
⋮	⋮	⋮	⋮
Πληθυσμός r	x_r	$n_r - x_r$	n_r

Μπορεί να δειχθεί ότι η στατιστική συνάρτηση ελέγχου είναι η

$$U = \sum_{i=1}^{r} \frac{(x_i - n_i p)^2}{n_i p(1-p)}$$

όπου $p = (1/n)\sum_{i=1}^{r} x_i$ και $n = n_1 + n_2 + \ldots + n_r$, η οποία ακολουθεί προσεγγιστικά την κατανομή χ^2_{r-1} όταν η H_0 είναι αληθής. Ουσιαστικά η U είναι το άθροισμα των τετραγώνων r τυποποιημένων διωνυμικών κατανομών με κοινή πιθανότητα επιτυχίας p η οποία εκτιμάται θεωρώντας ότι όλοι οι πληθυσμοί είναι ένας. Η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η

$\{u : u \geq \chi^2_{r-1;\,a}\}$ όπου u είναι η παρατηρούμενη τιμή της U. Η $p-value$ του ελέγχου είναι ίση με $P(\chi^2_{r-1} \geq u)$.

Ο προσεγγιστικός έλεγχος για την ισότητα *r* αναλογιών με ανεξάρτητα δείγματα γίνεται με τη βοήθεια της συνάρτησης `prop.test` (ή και με `chisq.test`). Η βασική σύνταξη της συνάρτησης `prop.test` είναι η ακόλουθη:

```
prop.test(x, n, correct=TRUE or FALSE)
x: διάνυσμα με τον αριθμό των επιτυχιών
n: διάνυσμα με τον αριθμό των δοκιμών
correct: υπολογισμός p-value με διόρθωση (TRUE) ή χωρίς διόρθωση
    (FALSE) συνέχειας
```

Παράδειγμα 6.35 (prop.test για *r* πληθυσμούς Bernoulli)

Σε τέσσερα νοσοκομεία που βρίσκονται σε διαφορετικές περιοχές της χώρας και για ένα διάστημα 6 μηνών καταγράφηκε ο αριθμός των διαγνώσεων καρκίνου του πνεύμονα (397 καταγραφές) και ο αριθμός των ασθενών που κάπνιζαν. Τα αποτελέσματα παρουσιάζονται στον ακόλουθο πίνακα.

	Καπνιστές	Μη καπνιστές	Σύνολο
Νοσοκομείο 1	83	3	86
Νοσοκομείο 2	90	3	93
Νοσοκομείο 3	129	7	136
Νοσοκομείο 4	70	12	82

Θέλουμε να ελέγξουμε αν το ποσοστό των καπνιζόντων ήταν το ίδιο και στα 4 νοσοκομεία. Ο έλεγχος γίνεται με R ως ακολούθως

```
> smokers  <- c( 83, 90, 129, 70 )
> patients <- c( 86, 93, 136, 82 )
> prop.test(smokers, patients)

4-sample test for equality of proportions without continuity correction

data:  smokers out of patients
X-squared = 12.6004, df = 3, p-value = 0.005585
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.9651163 0.9677419 0.9485294 0.8536585
```

Επιβεβαίωση της *p-value*:

```
> p <- sum(smokers)/sum(patients)
> numer <-(smokers-patients*p)^2
> denom <-patients*p*(1-p)
> U <-sum(numer/denom);U
[1] 12.60041
> df <- length(smokers)-1
> cat("Επιβεβαίωση: p-value =", 1-pchisq(U,df),"\n")
Επιβεβαίωση: p-value = 0.005585477
```

Επομένως το συμπέρασμα είναι ότι τα ποσοστά διαφέρουν στα 4 νοσοκομεία. ∎

6.5 Συμπερασματολογία για *k* δείγματα

6.5.1 Έλεγχος ισότητας *k* μέσων – Anova κατά ένα παράγοντα

Στην πράξη ερχόμαστε συχνά με το πρόβλημα του ελέγχου της ισότητας των μέσων k ανεξάρτητων δειγμάτων. Συνήθως τα k δείγματα αντιστοιχούν σε k διαφορετικά επίπεδα (**στάθμες**) μιας κατηγορικής μεταβλητής (παράγοντας). Για παράδειγμα, ένας αγρότης θα ήθελε να γνωρίζει ποιο λίπασμα, επί συνόλου τριών διαφορετικών λιπασμάτων, αποδίδει τη μεγαλύτερη παραγωγή σίτου σε Kg ανά στρέμμα. Ένας απλός τρόπος για να αντιμετωπιστεί το ερώτημα του αγρότη είναι ο ακόλουθος: Θα έπρεπε να είχαμε αρχικά στη διάθεσή μας n διαφορετικά αγροτεμάχια σε μια περιοχή με την ίδια μορφοποίησης του εδάφους. Στη συνέχεια, με ένα τυχαίο τρόπο, θα επιλέγαμε n_i ($i=1,2,3$) από αυτά όπου θα εφαρμόζαμε το λίπασμα i ($n_1+n_2+n_3=n$) και θα καταγράφαμε την απόδοση X_{ij} ($i=1,2,3,\ \ j=1,2,...,n_i$) ανά στρέμμα του j αγροτεμαχίου. Αν συμβολίσουμε με μ_i ($i=1,2,3$) τη μέση απόδοση ανά στρέμμα με τη χρήση του i λιπάσματος θα μας ενδιέφερε ο έλεγχος της ισότητας των τριών μέσων, δηλαδή ο έλεγχος

$$H_0: \mu_1=\mu_2=\mu_3 \quad - \quad H_1: \text{Όχι } H_0$$

του οποίου η μηδενική υπόθεση δηλώνει ότι ο τύπος του λιπάσματος δεν έχει επίδραση στη μέση στρεμματική απόδοση σε Kg

Στα πλαίσια της ανάλυσης διακύμανσης κατά ένα παράγοντα με σταθερές επιδράσεις (one-way anova with fixed effects) περιγράφουμε τις παρατηρήσεις μας με το γραμμικό μοντέλο

$$X_{ij}=\mu_i+\varepsilon_{ij},\quad i=1,2,...,k,\ \ j=1,2,...,n_i$$

όπου μ_i είναι η μέση τιμή που αντιστοιχεί στην i στάθμη και ε_{ij} είναι το τυχαίο σφάλμα που αντιστοιχεί στην παρατήρηση X_{ij}. Σε αυτό το μοντέλο μας ενδιαφέρει ο έλεγχος υπόθεσης

$$H_0: \mu_1=\mu_2=\ldots=\mu_k \quad - \quad H_1: \text{Όχι } H_0.$$

Ένα ισοδύναμο μοντέλο είναι το

$$X_{ij}=\mu+\tau_i+\varepsilon_{ij},\quad i=1,2,...,k,\quad j=1,2,...,n_i$$

όπου το μ_i έχει αντικατασταθεί με το $\mu+\tau_i$, δηλαδή $\tau_i=\mu-\mu_i$ ($i=1,2,...,k$). Η παράμετρος τ_i αναφέρεται ως επίδραση της αγωγής i (treatment effect). Ωστόσο το παραπάνω μοντέλο περιέχει $k+1$ παραμέτρους (overparameterized model) για να περιγράψει τους k μέσους, και επομένως επιβάλλεται συνήθως ο περιορισμός $\sum_{i=1}^{k} n_i\tau_i=0$. Ο έλεγχος της ισότητας των μέσων σύμφωνα με το παραπάνω μοντέλο ανάγεται στον

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \quad - \quad H_1 : \text{Όχι } H_0 .$$

Τα δεδομένα μας στην ανάλυση διακύμανσης κατά ένα παράγοντα παρουσιάζονται συνήθως με την ακόλουθη δομή.

Επίδραση					Σύνολο	Μέσοι
1	X_{11}	X_{12}	...	X_{1n_1}	$X_{1\bullet}$	$\overline{X}_{1\bullet}$
2	X_{21}	X_{22}	...	X_{2n_2}	$X_{2\bullet}$	$\overline{X}_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
k	X_{k1}	X_{k2}	...	X_{kn_k}	$X_{k\bullet}$	$\overline{X}_{k\bullet}$
					$X_{\bullet\bullet}$	$\overline{X}_{\bullet\bullet}$

Για την εκτέλεση του ελέγχου της ισότητας των μέσων κατασκευάζεται ο ακόλουθος πίνακας ANOVA ($n_1 + n_2 + \ldots + n_k = n$)

Πίνακας ANOVA

Προέλευση μεταβλητότητας (Source)	Βαθμοί ελευθερίας (Df)	Άθροισμα τετραγώνων (SS)	Μέσο τετράγωνο (MS)	Τιμή της F (F)
Αγωγή (Treatment)	$k-1$	$SS_{Treatment} = \sum_{i=1}^{k} n_i (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2$	$MS_{Treatment} = \frac{SS_{Treatment}}{k-1}$	$F = \frac{MS_{Treatment}}{MS_{Error}}$
Σφάλμα (Error)	$n-k$	$SS_{Error} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\bullet})^2$	$MS_{Error} = \frac{SS_{Error}}{n-k}$	
Σύνολο	$n-1$	$SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{\bullet\bullet})^2$		

Όταν η μηδενική υπόθεση είναι αληθής έχουμε ότι

$$F = \frac{MS_{Treatment}}{MS_{Error}} \sim F_{k-1,n-k} .$$

Αν συμβολίσουμε με *f* την παρατηρούμενη τιμή της *F*, η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η $\{f : f \geq F_{k-1,n-k;a}\}$. Η $p-value$ του ελέγχου είναι ίση με $P(F \geq f)$.

Για να ισχύουν όλα τα παραπάνω θα πρέπει οι παρατηρήσεις X_{ij} να είναι ανεξάρτητες τ.μ. και $X_{ij} \sim N(\mu_i, \sigma^2)$. Ισοδύναμα, τα σφάλματα ε_{ij} πρέπει να είναι ανεξάρτητες τ.μ. και $\varepsilon_{ij} \sim N(0, \sigma^2)$.

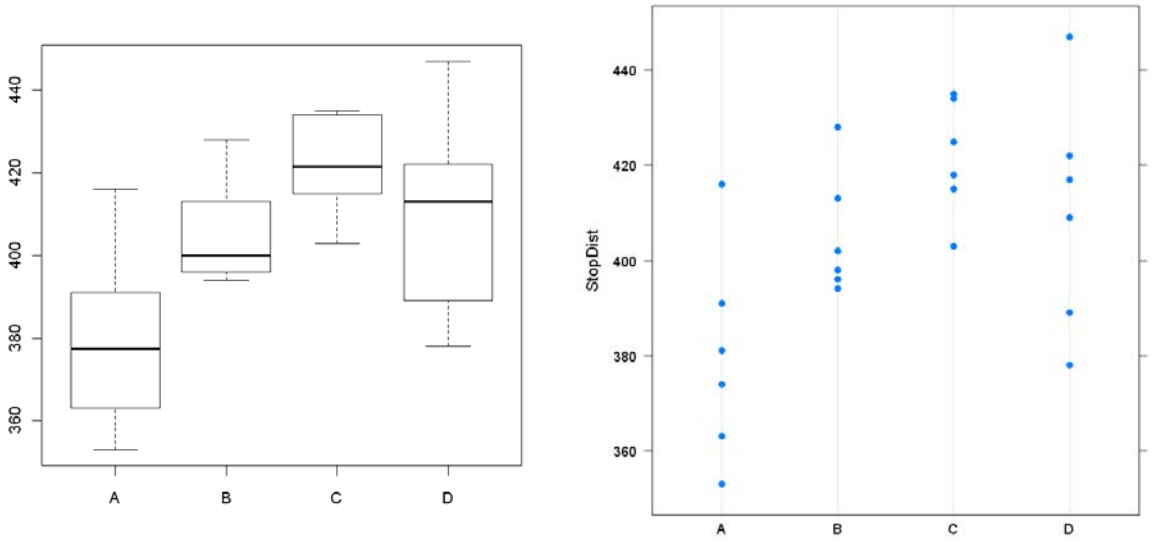
Για την εμφάνιση του πίνακα ANOVA με το R χρησιμοποιείται η συνάρτηση `aov`. Η βασική σύνταξη της συνάρτησης `aov` είναι η ακόλουθη

```
summary(aov(x~τ))
x: η μεταβλητή απόκρισης
τ: η μεταβλητή (παράγοντας, factor) που περιέχει τις στάθμες
```

Παράδειγμα 6.36 (summary(aov))

Το πλαίσιο δεδομένων `Tire` του πακέτου `PASWR` περιέχει 24 μετρήσεις που αναφέρονται στην απόσταση (σε ft) που διανύει ένα συγκεκριμένο αυτοκίνητο (με τον ίδιο πάντα οδηγό) από τη στιγμή που θα πατηθεί το φρένο έως ότου ακινητοποιηθεί πλήρως, όταν η ταχύτητά του είναι 60 μίλια ανά ώρα (μεταβλητή `StopDist`). Υπάρχουν 4 σετ των 6 μετρήσεων, όπου στο κάθε σετ χρησιμοποιήθηκαν ελαστικά με διαφορετικά πέλματα (μεταβλητή `tire` που είναι παράγοντας (factor) με τιμές A, B, C, D). Για να διαπιστώσουμε αν είναι ίδιες οι μέσες αποστάσεις που χρειάζεται για να σταματήσει το αυτοκίνητο με τα τέσσερα διαφορετικά πέλματα, εργαζόμαστε με το R ως εξής:

```
> library(PASWR)
> attach(Tire);names(Tire)
[1] "StopDist" "tire"
> plot(tire, StopDist)
>dotplot(StopDist~tire)
```



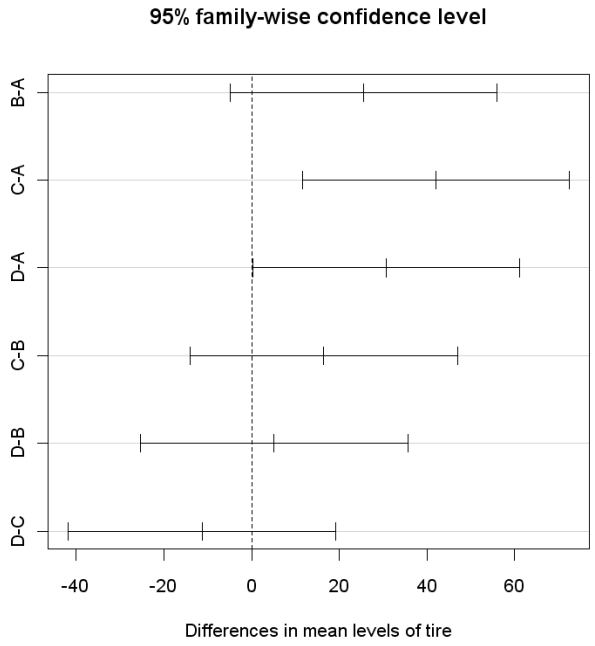
Από τα παραπάνω γράφημα αντιλαμβανόμαστε ότι οι μέσες αποστάσεις φρεναρίσματος πρέπει να διαφέρουν, ειδικά στους τύπους ελαστικών A και C. Ο πίνακας ANOVA προκύπτει ως εξής

```
> model <- aov(StopDist~tire)
> summary(model)
            Df Sum Sq Mean Sq F value   Pr(>F)
tire         3 5673.1 1891.04  5.3278 0.007316 **
Residuals   20 7098.8  354.94
```

Ο πίνακας ANOVA επιβεβαιώνει ότι οι 4 μέσες αποστάσεις φρεναρίσματος δεν είναι ίσες. Ένας τρόπος για να διαπιστωθεί ποιες μέσες τιμές διαφέρουν (όταν τις παίρνουμε ανά δύο) είναι να κατασκευαστεί ένα πολλαπλό διάστημα εμπιστοσύνης με (συνολικό) συντελεστή εμπιστοσύνης $1-a$

για τις διαφορές των μέσων τιμών ανά δύο (υπάρχουν $k(k-1)/2$ τέτοιες διαφορές) και να εξεταστεί ποια από τα διαστήματα περιέχουν το 0 (αν το περιέχουν τότε δεν διαφέρουν οι αντίστοιχες μέσες τιμές). Οι πολλαπλές συγκρίσεις των μέσων με τη μέθοδο του Tukey με το R προκύπτουν ως ακολούθως.

```
> TukeyHSD(model, conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = StopDist ~ tire)
$tire
          diff         lwr      upr     p adj
B-A  25.500000  -4.9446409 55.94464 0.1213153
C-A  42.000000  11.5553591 72.44464 0.0049515
D-A  30.666667   0.2220258 61.11131 0.0479540
C-B  16.500000 -13.9446409 46.94464 0.4464584
D-B   5.166667 -25.2779742 35.61131 0.9637307
D-C -11.333333 -41.7779742 19.11131 0.7273681
> plot(TukeyHSD(model))
```



Το σίγουρο είναι ότι η διαφορά $\mu_C - \mu_A$ (που είναι ίση με 42) είναι σημαντική και επομένως οι μέσες τιμές μ_A, μ_C διαφέρουν.

6.5.2 Έλεγχος Kruskal-Wallis

Στην προηγούμενη παράγραφο στα πλαίσια της ανάλυσης διακύμανσης κατά ένα παράγοντα περιγράψαμε ένα έλεγχο για την ισότητα των μέσων k ανεξάρτητων τυχαίων δειγμάτων υποθέτοντας ότι όλα τα δείγματα προέρχονται από πληθυσμούς που κατανέμονται κανονικά με ίσες διακυμάνσεις.

Στον έλεγχο υπόθεσης των Kruskal-Wallis γίνεται μόνο η υπόθεση ότι οι *k* πληθυσμοί έχουν συναρτήσεις κατανομής F_1, F_2, ..., F_k που προέρχονται από την ίδια γενική οικογένεια κατανομών F (όχι αναγκαστικά την κανονική κατανομή) οι οποίοι μπορούν να διαφέρουν μόνο ως προς την παράμετρο θέσης, αλλά όχι και ως προς την παράμετρο κλίμακας (διακύμανση). Αν και ορισμένοι συγγραφείς υποστηρίζουν ότι ο έλεγχος των Kruskal-Wallis μπορεί να εφαρμοστεί ακόμη και αν οι διακυμάνσεις των πληθυσμών δεν είναι ίσες, θα πρέπει να αναφέρουμε ότι σε αυτές τις περιπτώσεις υπάρχουν άλλοι πιο αξιόπιστοι έλεγχοι (Rust-Flinger τροποποίηση του Kruskal-Wallis ελέγχου).
Χρησιμοποιώντας τους συμβολισμούς της προηγούμενης παραγράφου θα ασχοληθούμε εδώ με τον έλεγχο (location model)

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k \quad - \quad H_1 : \text{Όχι } H_0 .$$

(συνήθως τα θ_i είναι οι διάμεσοι των *k* δειγμάτων). Η μηδενική υπόθεση ισχυρίζεται ότι οι *k* πληθυσμοί έχουν την ίδια κατανομή. Προκειμένου να εφαρμοστεί ο έλεγχος όλες οι $n_1 + n_2 + \ldots + n_k = n$ παρατηρήσεις ενώνονται σε ένα δείγμα και βαθμολογούνται από το 1 έως το n. Στη συνέχεια υπολογίζονται τα αθροίσματα R_i, $i = 1,2,...,k,$ των βαθμών που αντιστοιχούν στις παρατηρήσεις των *k* δειγμάτων. Η στατιστική συνάρτηση ελέγχου είναι η

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) .$$

Μεγάλες τιμές της *H* οδηγούν σε απόρριψη της μηδενικής υπόθεσης. Η κατανομή της *H* για μεγάλα δείγματα προσεγγίζεται από την κατανομή χ^2_{k-1} (η προσέγγιση θεωρείται ότι είναι καλή όταν $\min\{n_1, n_2, ..., n_k\} \geq 5$) και επομένως η προσεγγιστική $p-value$ του ελέγχου είναι ίση με $P(\chi^2_{k-1} \geq h)$ όπου h είναι η παρατηρούμενη τιμή της *H*. Αν υπάρχουν δεσμοί τότε αντί της *H* χρησιμοποιούμε τη στατιστική συνάρτηση ελέγχου

$$H_t = \frac{H}{1 - \left(\sum_{i=1}^{r} t_i (t_i^2 - 1) / n(n^2 - 1) \right)} .$$

Στον παραπάνω τύπο το r δηλώνει το πλήθος των διαφορετικών βαθμών, και το t_i δηλώνει πόσες φορές εμφανίζεται κάθε διαφορετικός βαθμός $(1 \leq i \leq r)$. Παρατηρούμε ότι όταν δεν υπάρχουν δεσμοί ($t_i = 1$) τότε η H_t ανάγεται στην H.

Μπορεί να αποδειχθεί ότι ο έλεγχος Kruskal-Wallis για δύο δείγματα είναι ισοδύναμος με τον έλεγχο Wilcoxon Rank-Sum (δίπλευρη εναλλακτική). Αν και θεωρητικά οι *k* πληθυσμοί περιγράφονται από συνεχείς κατανομές, ο έλεγχος Kruskal-Wallis μπορεί να εφαρμοστεί και σε διατάξιμα δεδομένα.

Για τον παραπάνω έλεγχο στο R χρησιμοποιείται η συνάρτηση `kruskal.test`. Η βασική σύνταξη της συνάρτησης `kruskal.test` είναι η ακόλουθη

```
kruskal.test(list(x1, x2,...,xk))
xi: διάνυσμα που περιέχει τις παρατηρήσεις του i δείγματος

ή

kruskal.test(x, t)
x: η μεταβλητή απόκρισης
t: η μεταβλητή που περιέχει τις στάθμες
```

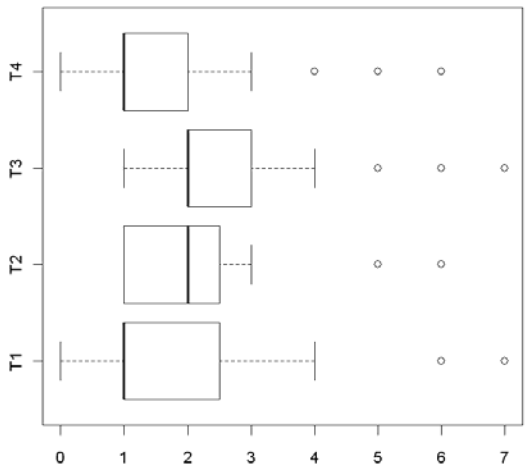
Παράδειγμα 6.37 (kruskal.test)

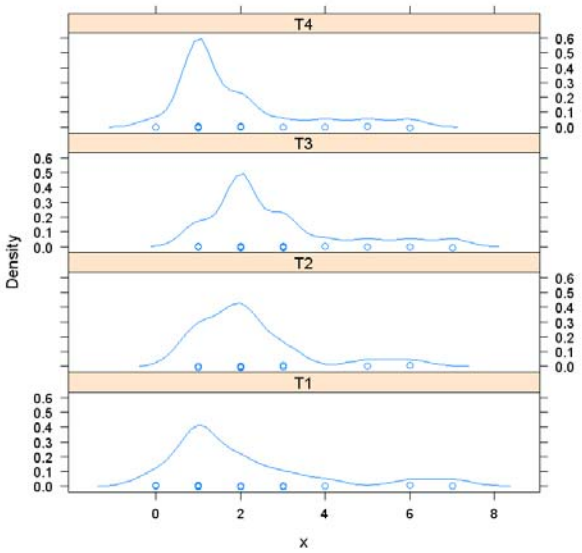
Ένας γυμναστής ενδιαφέρεται να αξιολογήσει τέσσερις διαφορετικές τεχνικές ρίψης ελεύθερων βολών στο μπάσκετ. Χωρίζει με τυχαίο τρόπο ένα σύνολο 80 παιδιών σε τέσσερις ομάδες των 20 παιδιών και προπονεί κάθε ομάδα επί 2 μήνες σε διαφορετική τεχνική ρίψης ελεύθερων βολών. Στη συνέχεια κάθε παιδί ρίχνει 10 ελεύθερες βολές και ο προπονητής καταγράφει τον αριθμό των επιτυχημένων βολών που δίνονται στο ακόλουθο πλαίσιο.

Τεχνική 1	6	1	2	0	0	1	1	3	1	2	1	2	4	2	1	1	1	3	7	1
Τεχνική 2	3	2	1	2	1	6	2	1	1	2	1	1	2	3	2	2	3	2	5	2
Τεχνική 3	2	1	2	3	2	2	4	3	2	3	2	5	1	1	3	7	6	2	2	2
Τεχνική 4	2	1	1	3	1	2	1	6	1	1	0	1	1	1	1	2	2	1	5	4

Εκτελούμε αρχικά μια περιγραφική στατιστική ανάλυση των δεδομένων.

```
> x1 <- c(6,1,2,0,0,1,1,3,1,2,1,2,4,2,1,1,1,3,7,1)
> x2 <- c(3,2,1,2,1,6,2,1,1,2,1,1,2,3,2,2,3,2,5,2)
> x3 <- c(2,1,2,3,2,2,4,3,2,3,2,5,1,1,3,7,6,2,2,2)
> x4 <- c(2,1,1,3,1,2,1,6,1,1,0,1,1,1,1,2,2,1,5,4)
> x <- c(x1, x2, x3, x4)
> d <- rep(c("T1","T2","T3","T4"), each=20);
> t <- factor(d)
> plot(t, x, horizontal=T)
> library(lattice)
> densityplot(~x|t, layout=c(1,4))
```





Από τα παραπάνω σχήματα μπορούμε να υποθέσουμε ότι οι 4 πληθυσμοί προέρχονται από την ίδια γενική οικογένεια κατανομών. Εκτελώντας τον έλεγχο Kruskal-Wallis παίρνουμε τα παρακάτω αποτελέσματα.

```
> kruskal.test(list(x1,x2,x3,x4))

        Kruskal-Wallis rank sum test

data:  list(x1, x2, x3, x4)
Kruskal-Wallis chi-squared = 7.8654, df = 3, p-value = 0.04888
```

Έτσι προκύπτει ότι οι τέσσερις μέθοδοι διαφέρουν σε επίπεδο σημαντικότητας 0.05.

6.5.3 Έλεγχος Levene για την ισότητα *k* διακυμάνσεων

Σε αρκετές περιπτώσεις μας ενδιαφέρει να ελέγξουμε την ισότητα των διακυμάνσεων k πληθυσμών, δηλαδή μας ενδιαφέρει ο έλεγχος

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 \quad - \quad H_1 : \text{Όχι } H_0 .$$

Μια συνήθης επιλογή για τον παραπάνω έλεγχο είναι ο έλεγχος του Levene ο οποίος δεν απαιτεί κανονικότητα των k πληθυσμών (είναι λιγότερο ευαίσθητος έλεγχος από τον έλεγχο του Bartlett σε αποκλίσεις από την κανονικότητα). Χρησιμοποιώντας τους συμβολισμούς της Παραγράφου 6.5.1 έχουμε ότι η στατιστική συνάρτηση ελέγχου δίνεται από τον τύπο

$$W = \frac{(n-k)\sum_{i=1}^{k} n_i (\overline{Z}_{i\bullet} - \overline{Z}_{\bullet\bullet})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} (Z_{ij} - \overline{Z}_{i\bullet})^2}$$

όπου $Z_{ij} = | X_{ij} - m_i |$ $(i = 1,2,...,k,\ \ j = 1,2,...,n_i)$. Η ποσότητα m_i μπορεί να είναι είτε (α) ο μέσος του i δείγματος $(m_i = \overline{X}_{i\bullet})$, είτε (β) η διάμεσος του i δείγματος, είτε (γ) ο περικομμένος μέσος

(trimmed mean) του *i* δείγματος. Σημειώνουμε ότι περικομμένος μέσος κατά 10% σημαίνει ότι ο δειγματικός μέσος υπολογίζεται αφού απαλειφθεί το άνω 10% και το κάτω 10% των παρατηρήσεών μας. Στην περίπτωση (α) έχουμε το κλασικό τεστ του Levene, στην περίπτωση (β) έχουμε το Brown-Forsythe Levene τεστ, και στην περίπτωση (γ) το τεστ του Levene με περικομμένο μέσο. Το Brown-Forsythe Levene τεστ θεωρείται μια καλή επιλογή αφού το τεστ είναι ανθεκτικό έναντι αρκετών τύπων μη κανονικών κατανομών διατηρώντας ταυτόχρονα αυξημένη ισχύ.

Όταν η μηδενική υπόθεση είναι αληθής η στατιστική συνάρτηση W ακολουθεί την κατανομή $F_{k-1,n-k}$. Μεγάλες τιμές της W οδηγούν σε απόρριψη της H_0 και επομένως η κρίσιμη περιοχή του ελέγχου σε επίπεδο σημαντικότητας a είναι η $\{w: w \geq F_{k-1,n-k;a}\}$ όπου w είναι η παρατηρούμενη τιμή της W. Η $p-value$ του ελέγχου είναι ίση με $P(F_{k-1,n-k} \geq w)$.

Στο R χρησιμοποιείται η συνάρτηση `levene.test` του πακέτου `lawstat` για τον έλεγχο της ισότητας των διακυμάνσεων *k* πληθυσμών (δείτε επίσης Παράγραφο 6.3.4). Η βασική σύνταξη της συνάρτησης `levene.test` είναι η ακόλουθη

```
levene.test(x, t, location=c("median", "mean", "trim.mean")
x: η μεταβλητή απόκρισης
t: η μεταβλητή που περιέχει τις στάθμες
location: δηλώνει το σημείο γύρω από το οποίο υπολογίζονται οι α-
    ποκλίσεις των δεδομένων μας.
```

Παράδειγμα 6.38 (levene.test)

Ανακαλούμε τα δεδομένα του Παραδείγματος 6.36 για να διαπιστώσουμε αν οι διακυμάνσεις των αποστάσεων με τα 4 διαφορετικά πέλματα είναι ίσες. Άλλωστε είναι απαραίτητο να είναι ίσες αφού σε διαφορετική περίπτωση θα ήμασταν ιδιαίτερα προβληματισμένοι για την αξιοπιστία του συμπεράσματος του Παραδείγματος 6.36 (ότι δηλαδή οι 4 μέσες αποστάσεις φρεναρίσματος δεν είναι ίσες). Εργαζόμαστε με το R ως εξής

```
> library(PASWR); > library(lawstat)
> attach(Tire);names(Tire)
-------------------------------------------------------------------------
> levene.test(StopDist, tire, location="mean")

classical Levene's test based on the absolute deviations from the mean
        ( none not applied because the location is not set to median )

data:  StopDist
Test Statistic = 0.9896, p-value = 0.4178
-------------------------------------------------------------------------
> levene.test(StopDist, tire, location="median")

modified robust Brown-Forsythe Levene-type test based on the absolute
deviations from the median

data:  StopDist
Test Statistic = 0.9789, p-value = 0.4224
```

Επομένως δεχόμαστε την ισότητα των διακυμάνσεων των τεσσάρων πληθυσμών.

6.6 Γραμμική παλινδρόμηση

Το βασικό μοντέλο που θα μας απασχολήσει στην παρούσα παράγραφο είναι το κλασικό μοντέλο γραμμικής παλινδρόμησης

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

όπου Y είναι η μεταβλητή απόκρισης (response), X_i $(1 \le i \le p)$ είναι οι ερμηνευτικές (predictors) μεταβλητές και για το σφάλμα ε έχουμε ότι $\varepsilon \sim N(0, \sigma^2)$. Στον παρακάτω πίνακα δίνουμε μερικές βασικές μορφές μοντέλων και η αντίστοιχη δήλωσή τους στο *R*.

Μοντέλο	Περιγραφή στο R
$Y = \beta_0 + \beta_1 x + \varepsilon$	$y \sim x$ ή $y \sim 1 + x$
$Y = \beta + \varepsilon$	$y \sim 1$
$Y = \beta x + \varepsilon$	$y \sim 0 + x$ ή $y \sim -1 + x$ ή $y \sim x - 1$
$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$	$y \sim 1 + x + I(x\text{^}2)$
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	$y \sim x1 + x2$
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$	$y \sim x1 + x2 + x1 : x2$ ή $y \sim x1 * x2$

6.6.1 Απλή γραμμική παλινδρόμηση

Στην παρούσα παράγραφο θα εξετάσουμε το απλό γραμμικό πρότυπο

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

μέσω ενός συγκεκριμένου παραδείγματος. Θεωρούμε τα δεδομένα του ακόλουθου πίνακα

Θερμοκρασία x_i	140	145	150	155	160	165
Απόδοση y_i	14.1	14.4	15.1	18.1	18.3	20.5

που αναφέρονται στην απόδοση μιας διεργασίας σε έξι διαφορετικές θερμοκρασίες. Η προσαρμοσμένη ευθεία γραμμικής παλινδρόμησης είναι η

$$\hat{y} = -23.946 + 0.2668x\,.$$

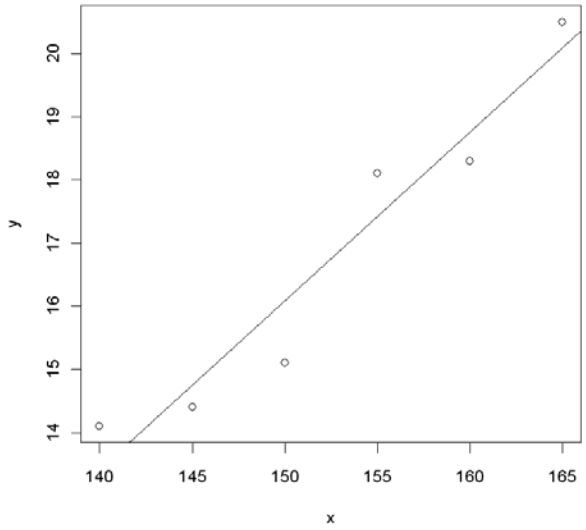
Η ανάλυση των γραμμικών μοντέλων στο R γίνεται με τη συνάρτηση `lm`. Έτσι μπορούμε να επιβεβαιώσουμε το παραπάνω αποτέλεσμα ως ακολούθως

```
> x <- c(140, 145, 150, 155, 160, 165)
> y <- c(14.1, 14.4, 15.1, 18.1, 18.3, 20.5)
> lm(y~x)

Call:
```

```
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   -23.9457       0.2669
```

Για την οπτική παρουσίαση του αποτελέσματος έχουμε

```
> v <- lm(y~x); plot(x,y); abline(v)
```



Μια βασική σχέση που ισχύει στη γραμμική παλινδρόμηση είναι η ακόλουθη

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

που δηλώνεται ως $SST = SSR + SSE$.

Περαιτέρω ανάλυση του μοντέλου προκύπτει με τη συνάρτηση `summary`. Για παράδειγμα

```
> summary(v)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6
 0.6857 -0.3486 -0.9829  0.6829 -0.4514  0.4143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.94571    5.65568  -4.234  0.01333 *
x             0.26686    0.03703   7.207  0.00197 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7745 on 4 degrees of freedom
Multiple R-squared: 0.9285,     Adjusted R-squared: 0.9106
F-statistic: 51.94 on 1 and 4 DF,  p-value: 0.001965
```

Η πρώτη γραμμή της ανάλυσης δίνει τα υπόλοιπα (residuals) $e_i = y_i - \hat{y}_i$. Στη συνέχεια δίνονται για τους εκτιμητές ελαχίστων τετραγώνων των β_0 (`Intercept`) και β_1 (`x`) οι εκτιμήσεις των (στήλη `Estimate`), εκτιμήσεις των τυπικών σφαλμάτων (στήλη `Std. Error`) και τα αποτελέσματα των ελέγχων

$$H_0 : \beta_0 = 0 \quad - \quad H_1 : \beta_0 \neq 0 \quad H_0 : \beta_1 = 0 \quad - \quad H_1 : \beta_1 \neq 0 .$$

Για τους ελέγχους δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (στήλη `t value` = Estimate / Std. Error) η οποία ακολουθεί την κατανομή t_{n-2}, καθώς και η $p-value$ του ελέγχου (στήλη `Pr(>|t|)`). Στη συνέχεια δίνεται εκτίμηση για την τυπική απόκλιση σ των σφαλμάτων (`Residual standard error`) με τη βοήθεια της σχέσης $\hat{\sigma}^2 = s^2 = \sum e_i^2 / (n-2) = SSE/(n-2)$.

Ακολουθεί εκτίμηση του συντελεστή προσδιορισμού $r^2 = SSR/SST$ (`Multiple R-squared`) καθώς επίσης και μια προσαρμοσμένη τιμή του (`Adjusted R-squared`). Στο τέλος δίνεται το αποτέλεσμα του ελέγχου

$$H_0 : \beta_1 = 0 \quad - \quad H_1 : \beta_1 \neq 0 .$$

(έλεγχος σημαντικότητας της παλινδρόμησης). Για τον έλεγχο δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (`F-statistic` = $SSR/[SSE/(n-2)]$) η οποία ακολουθεί την κατανομή $F_{1,n-2}$ καθώς και η $p-value$ του ελέγχου. Στα πλαίσια της απλής γραμμικής παλινδρόμησης παρατηρούμε ότι ο παραπάνω έλεγχος μπορεί να εκτελεστεί με δύο ισοδύναμους τρόπους.

Ο πίνακας της ανάλυσης διασποράς προκύπτει με τη συνάρτηση `anova`. Για παράδειγμα

```
> anova(v)
Analysis of Variance Table


Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x          1 31.1556 31.1556  51.938 0.001965 **
Residuals  4  2.3994  0.5999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ο παραπάνω πίνακας έχει τη μορφή

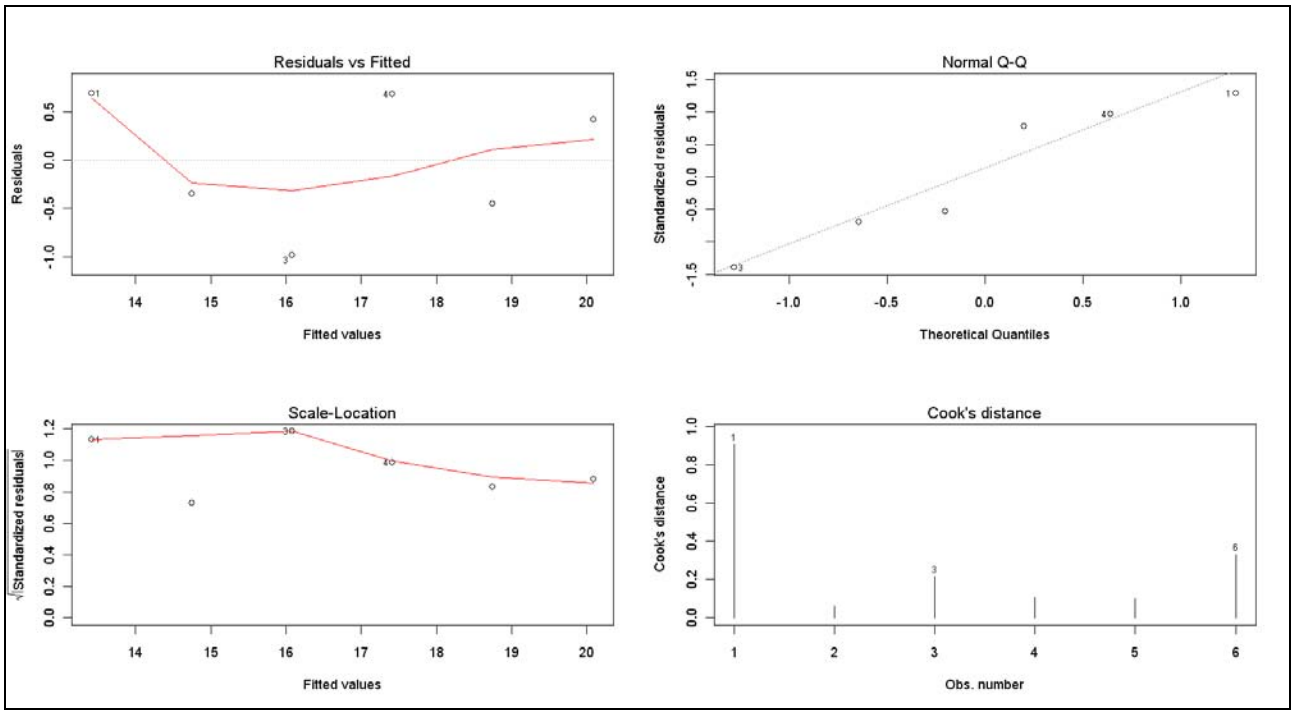
Source	DF	SS	MS	F
Regression	1	SSR	$SSR/1$	SSR/s^2
Residual Error	$n-2$	SSE	$s^2 = SSE/(n-2)$	

Για να πάρουμε διαστήματα εμπιστοσύνης μέσης και ατομικής πρόβλεψης για το Y χρησιμοποιούμε τη συνάρτηση `predict`. Για παράδειγμα

```
> predict(v, newdata=data.frame(x=153), interval="confidence",
+ level=0.95)            ### Μέση πρόβλεψη
       fit      lwr      upr
1 16.88343 16.00404 17.76282
> predict(v, newdata=data.frame(x=153), interval="prediction",
+ level=0.95) )          ### Ατομική πρόβλεψη
       fit      lwr      upr
1 16.88343 14.56020 19.20666
```

Με τη συνάρτηση `plot` λαμβάνουμε γραφικές παραστάσεις σχετικές με το μοντέλο μας. Για παράδειγμα

```
> par(mfrow=c(2,2))
> plot(v,which = 1:4)
```



Στο πρώτο διάγραμμα (Residuals vs Fitted) δίνεται ένα διάγραμμα διασποράς των σημείων $(\hat{y}_i, e_i)$, $1 \le i \le n$. Στην ιδανική περίπτωση το διάγραμμα διασποράς θα πρέπει να δίνει την εικόνα των άστρων του ουρανού τη νύχτα. Αν εμφανίζεται κάποια συγκεκριμένη δομή (καμπύλωση) τότε αμφιβάλουμε για την καταλληλότητα του γραμμικού μοντέλου για την περιγραφή των δεδομένων μας (το ίδιο διάγραμμα μπορεί να χρησιμοποιηθεί για να διαπιστώσουμε αν η διακύμανση των σφαλμάτων είναι σταθερή). Στο δεύτερο διάγραμμα (Normal Q-Q) δίνεται ένα κανονικό Q-Q διάγραμμα των κανονικοποιημένων υπολοίπων (standardized residuals) το οποίο μας βοηθά να ελέγξουμε αν τα σφάλματα κατανέμονται κανονικά. Το τρίτο διάγραμμα (Scale-location), που είναι παρόμοιο με το πρώτο, μπορεί να χρησιμοποιηθεί για να ανιχνεύσει αν η διακύμανση των σφαλμάτων είναι σταθερή. Το τέταρτο διάγραμμα (Cook's distance) χρησιμοποιείται για να αναγνωριστούν τα σημεία

που έχουν μεγάλη επιρροή στην διαμόρφωση της εκτίμησης των συντελεστών παλινδρόμησης (εδώ τα σημεία 1 και 6).

6.6.2 Πολλαπλή γραμμική παλινδρόμηση

Στην παρούσα παράγραφο θα εξετάσουμε το γραμμικό πρότυπο

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

($p = 3$) μέσω ενός συγκεκριμένου παραδείγματος. Θεωρούμε τα δεδομένα του ακόλουθου πίνακα

x_1	x_2	x_3	Y
12.0	35.4	77.6	85.7
11.0	34.2	83.0	93.6
14.0	29.9	71.6	81.0
21.0	32.5	75.3	75.2
27.0	39.8	80.4	83.1
19.0	26.6	69.5	71.7
22.0	37.2	68.0	79.9
17.0	29.2	65.0	71.4
30.0	35.8	87.2	80.1
29.0	33.0	76.8	67.8
5.0	31.0	66.3	88.2
25.0	28.6	85.1	78.9
23.0	37.7	78.9	88.7
19.0	30.8	78.2	81.8
15.0	33.7	73.0	84.0

όπου στη στήλη Y δίνονται τιμές ενός δείκτη ατμοσφαιρικής ρύπανσης σε 15 τυχαία επιλεγμένες ημέρες του καλοκαιριού, ενώ στις στήλες x_1, x_2 και x_3 δίνονται οι αντίστοιχες τιμές της ταχύτητας του ανέμου (Km/h), της θερμοκρασίας (^oC) και της σχετικής υγρασίας (%). Η προσαρμοσμένη ευθεία γραμμικής παλινδρόμησης είναι η

$$\hat{y} = 21.5 - 0.929 x_1 + 1.001 x_2 + 0.582 x_3 .$$

Το παραπάνω αποτέλεσμα επιβεβαιώνεται ως ακολούθως

```
> x1 <- c(12,11,14,21,27,19,22,17,30,29,5,25,23,19,15)
> x2 <- c(35.4,34.2,29.9,32.5,39.8,26.6,37.2,29.2,
+ 35.8,33,31,28.6,37.7,30.8,33.7)
> x3 <- c(77.6,83,71.6,75.3,80.4,69.5,68,65,87.2,76.8,
+ 66.3,85.1,78.9,78.2,73)
> y <- c(85.7,93.6,81,75.2,83.1,71.7,79.9,71.4,80.1,67.8,
+ 88.2,78.9,88.7,81.8,84)
> lm(y~x1+x2+x3)

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
    21.4790      -0.9290       1.0009       0.5824
```

Περαιτέρω ανάλυση του μοντέλου προκύπτει με τη συνάρτηση `summary`. Έτσι

```
> m <- lm(y~x1+x2+x3)
> summary(m)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2573 -0.7993  0.2104  1.6626  4.9027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.4790    10.5795   2.030  0.06722 .
x1           -0.9290     0.1331  -6.978 2.34e-05 ***
x2            1.0009     0.2343   4.271  0.00132 **
x3            0.5824     0.1418   4.107  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.029 on 11 degrees of freedom
Multiple R-squared: 0.8554,     Adjusted R-squared: 0.816
F-statistic:  21.7 on 3 and 11 DF,  p-value: 6.345e-05
```

Η πρώτη γραμμή της ανάλυσης δίνει ορισμένα περιγραφικά μέτρα των υπολοίπων $e_i = y_i - \hat{y}_i$. Στη συνέχεια δίνονται οι εκτιμητές ελαχίστων τετραγώνων των παραμέτρων β_j $(0 \le i \le 3)$, τα τυπικά σφάλματα των εκτιμητών, και τα αποτελέσματα των ελέγχων

$$H_0 : \beta_j = 0 \quad - \quad H_1 : \beta_j \neq 0 \quad (j = 0, 1, 2, 3).$$

Για τους ελέγχους δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (στήλη `t value` = Estimate / Std. Error) η οποία ακολουθεί την κατανομή t_{n-p-1} (εδώ $p = 3$), καθώς και η $p-value$ του ελέγχου (στήλη `Pr(>|t|)`). Στη συνέχεια δίνεται εκτίμηση για την τυπική απόκλιση σ των σφαλμάτων (`Residual standard error`) με τη βοήθεια της σχέσης $\hat{\sigma}^2 = s^2 = SSE/(n-p-1)$. Ακολουθεί εκτίμηση του συντελεστή προσδιορισμού $r^2 = SSR/SST$ (`Multiple R-squared`) καθώς επίσης και μια προσασμοσμένη τιμή του (`Adjusted R-squared`). Στο τέλος δίνεται το αποτέλεσμα του ελέγχου

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad - \quad H_1 : \text{Όχι } H_0$$

(έλεγχος σημαντικότητας της παλινδρόμησης). Για τον έλεγχο δίνεται η τιμή της στατιστικής συνάρτησης ελέγχου (`F-statistic` = $(SSR/p)/[SSE/(n-p-1)]$) η οποία ακολουθεί την κατανομή $F_{p,n-p-1}$ καθώς η $p-value$ του ελέγχου.

Διαστήματα εμπιστοσύνης για τις παραμέτρους β_j $(j = 0, 1, 2, 3)$ του μοντέλου προκύπτουν με τη συνάρτηση συνάρτηση `confint`.

```
> confint(m,level=0.95)
                 2.5 %     97.5 %
(Intercept) -1.8062657 44.7643265
x1          -1.2220477 -0.6359752
x2           0.4851061  1.5166120
x3           0.2703105  0.8945337
```

Για να πάρουμε διαστήματα εμπιστοσύνης μέσης και ατομικής πρόβλεψης για το Y χρησιμοποιούμε τη συνάρτηση `predict`. Για παράδειγμα

```
> predict(m, newdata=data.frame(x1=8,x2=35,x3=83),
+  interval="confidence",level=0.99)   )   ###Μέση πρόβλεψη
       fit      lwr      upr
1 97.41804 90.14913 104.6870
> predict(m, newdata=data.frame(x1=8,x2=35,x3=83),
+ interval="prediction", level=0.99)       ###Ατομική πρόβλεψη
       fit      lwr      upr
1 97.41804 85.53045 109.3056
```

Για τη διερεύνηση του προβλήματος της πολυσυγγραμμικότητας μέσω των δεικτών VIF (variance inflation factor, παραγοντας διόγκωσης διακύμανσης) χρησιμοποιείται η συνάρτηση `vif` του πακέτου `DAAG`.

```
> library(DAAG)
> vif(m)
    x1     x2     x3
1.3584 1.1670 1.4077
```

Αν κάποιος δείκτης VIF πάρει τιμη μεγαλύτερη του 5 (ή 10 για άλλους συγγραφείς) τότε η αντίστοιχη μεταβλητή εμφανίζει πρόβλημα πολυσυγγραμμικότητας με τις υπόλοιπες ανεξάρτητες μεταβλητές.

Για τα πρότυπα γραμμικής παλινδρόμησης

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p + \varepsilon \quad \text{(full model, } f\text{)}$$

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon \text{ (restricted or reduced model, } r\text{)}$$

ο σχετικός έλεγχος υπόθεσης

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_{q+p} = 0 \quad - \quad H_1 : \text{Όχι } H_0$$

(απόρριψη της H_0 σημαίνει ότι οι μεταβλητές X_{q+1}, X_{q+2}, ..., X_p είναι ουσιαστικές στην εξήγηση της μεταβλητότητας του μοντέλου και, επομένως, πρέπει να προτιμηθεί το πλήρες πρότυπο) χρησιμοποιεί ως στατιστική συνάρτηση ελέγχου την

$$F = \frac{(SSR_f - SSR_r)/(p-q)}{SSE_f/(n-p-1)}$$

που ακολουθεί την κατανομή $F_{p-q,n-p-1}$ όταν η H_0 είναι ορθή. Ο παραπάνω έλεγχος στο R γίνεται με τη συνάρτηση `anova`. Για παράδειγμα

```
> f1 <- lm(y~x1)
> r1 <- lm(y~1)
> anova(r1,f1)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ x1
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     14 697.98
2     13 525.63  1    172.34 4.2624 0.05951 .
```

Επομένως σε επίπεδο σημαντικότητας $a = 0.10$ είναι απαραίτητη η εισαγωγή της ερμηνευτικής μεταβλητής x_1 στο μοντέλο που δεν περιέχει καμία μεταβλητή. Για την ανάγκη εισαγωξής και της x_2 έχουμε:

```
> f2 <- lm(y~x1+x2)
> r2 <- lm(y~x1)
> anova(r2,f2)
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     13 525.63
2     12 255.63  1       270 12.675 0.003922 **
```

Επομένως, στο μοντέλο που έχει ως μοναδική ερμηνευτική μεταβλητή τη x_1 είναι απαραίτητη η προσθήκη της ερμηνευτικής μεταβλητής x_2. Για την ανάγκη εισαγωξής και της x_3 έχουμε:

```
> f3 <- lm(y~x1+x2+x3)
> r3 <- lm(y~x1+x2)
> anova(r3,f3)
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     12 255.63
2     11 100.90  1    154.73 16.869 0.001738 **
```

Επομένως, στο μοντέλο που έχει ως ερμηνευτικές μεταβλητές τη x_1 και τη x_2 είναι απαραίτητη η προσθήκη της ερμηνευτικής μεταβλητής x_3.

Ο έλεγχος

$$H_0 : \beta_2 = \beta_3 = 0 \quad - \quad H_1 : \text{Όχι } H_0$$

εκτελείται ως ακολούθως

```
> f4 <- lm(y~x1+x2+x3)
> r4 <- lm(y~x1)
> anova(r4,f4)
Analysis of Variance Table
```

```
Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 525.63
2     11 100.90  2    424.74 23.153 0.0001142 ***
```

Για την εύρεση του μοντέλου που εφαρμόζει καλύτερα στα δεδομένα έχουν αναπτυχθεί διάφορες συναρτήσεις που χρησιμοποιούν το κριτήριο πληροφορίας του AIC. Το "καλύτερο" μοντέλο είναι αυτό με τη μικρότερη τιμή AIC. Με τη συνάρτηση `drop1` αφαιρούνται από το αρχικό μοντέλο μια μια οι μεταβλητές. Για παράδειγμα

```
> drop <- drop1(m)
> drop
Single term deletions

Model:
y ~ x1 + x2 + x3
       Df Sum of Sq    RSS    AIC
<none>              100.90 36.591
x1      1    446.60 547.50 59.960
x2      1    167.33 268.23 49.257
x3      1    154.73 255.63 48.535
```

Έτσι το μοντέλο και με τις τρεις μεταβλητές θεωρείται το καλύτερο αφού αν αφαιρεθεί οποιαδήποτε από τις μεταβλητές το AIC μεγαλώνει. Αν κάποια AIC ήταν μικρότερα από την τιμή 36.591 τότε θα απομακρύναμε εκείνη τη μεταβλητή που θα είχε το μικρότερο AIC, και θα συνεχίζαμε τη διαδικασία από την αρχή με το μοντέλο που δεν θα περιέχει τη συγκεκριμένη μεταβλητή (μια τέτοια διαδικασία οδηγεί σε επιλογή μεταβλητών σύμφωνα με τη μέθοδο Backward Elimination).

Με τη συνάρτηση `add1` προστίθενται στο μηδενικό μοντέλο μια μια οι μεταβλητές ξεκινώντας από το null model. Για παράδειγμα

```
> m0 <- lm(y~1)
> am0 <- add1(m0,scope=(~x1+x2+x3))
> am0
Single term additions

Model:
y ~ 1
       Df Sum of Sq    RSS    AIC
<none>              697.98 59.602
x1      1    172.34 525.63 57.348
x2      1    137.82 560.16 58.302
x3      1     55.22 642.76 60.366
```

Συμπεραίνουμε ότι είναι απαραίτητη η εισαγωγή της ερμηνευτικής μεταβλητής x_1 στο μοντέλο αφού μειώνεται το AIC. Συνεχίζοντας με τον ίδιο τρόπο παίρνουμε

```
> m1 <- lm(y~x1)
> am1 <- add1(m1,scope=(~x1+x2+x3))
> am1
Single term additions
```

```
Model:
y ~ x1
       Df Sum of Sq    RSS    AIC
<none>              525.63 57.348
x2      1    270.00 255.63 48.535
x3      1    257.40 268.23 49.257
```

Συμπεραίνουμε ότι είναι απαραίτητη η εισαγωγή της ερμηνευτικής μεταβλητής x_2 στο μοντέλο αφού μειώνεται το AIC. Συνεχίζοντας με τον ίδιο τρόπο παίρνουμε

```
> m2 <- lm(y~x1+x2)
> am2 <- add1(m2,scope=(~x1+x2+x3))
> am2
Single term additions

Model:
y ~ x1 + x2
       Df Sum of Sq    RSS    AIC
<none>              255.63 48.535
x3      1    154.73 100.90 36.591
```

Συμπεραίνουμε ότι είναι απαραίτητη η εισαγωγή της ερμηνευτικής μεταβλητής x_3 στο μοντέλο αφού μειώνεται το AIC, δηλαδή και οι τρεις μεταβλητές είναι απαραίτητες (μια τέτοια διαδικασία οδηγεί σε επιλογή μεταβλητών σύμφωνα με τη μέθοδο Forward Selection).

Εναλλακτικά μπορεί να χρησιμοποιηθεί η συνάρτηση `stepAIC` του πακέτου MASS η οποία ψάχνει να βρει το μοντέλο με το μικρότερο AIC. Το όρισμα `direction` δηλώνει την κατεύθυνση στην οποία θα κινηθεί η διαδικασία για την εύρεση του καλύτερου μοντέλου, ενώ το όρισμα `k` παίρνει συνήθως 2 τιμές. Την τιμή 2 και την τιμή log(n) όπου n είναι το μέγεθος μιας μεταβλητής. Θα εφαρμόσουμε τη συνάρτηση `stepAIC` σε ένα σύνολο δεδομένων με 6 ερμηνευτικές μεταβλητές.

```
> AIC <-read.table("AIC.txt", header=TRUE)
> attach(AIC)
> library(MASS) # For function stepAIC()
> m <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6)
> mod1 <- stepAIC(m, direction="both",scope=(~X6+X5+X4+X3+X2+X1), k=2)
Start:  AIC=179.51
Y ~ X1 + X2 + X3 + X4 + X5 + X6

       Df Sum of Sq    RSS    AIC
- X2    1     1.613 652.66 177.70
- X3    1     2.546 653.60 177.81
- X6    1     5.921 656.97 178.21
- X1    1     9.594 660.64 178.65
<none>              651.05 179.51
- X5    1    72.683 723.73 185.76
- X4    1   162.000 813.05 194.84

Step:  AIC=177.7
```

```
Y ~ X1 + X3 + X4 + X5 + X6

       Df Sum of Sq    RSS    AIC
- X6    1     5.693 658.36 176.38
- X1    1     9.875 662.54 176.87
- X3    1    10.554 663.22 176.95
<none>              652.66 177.70
+ X2    1     1.613 651.05 179.51
- X5    1    78.809 731.47 184.59
- X4    1   189.072 841.73 195.54

Step:  AIC=176.38
Y ~ X1 + X3 + X4 + X5

       Df Sum of Sq    RSS    AIC
- X3    1     6.833 665.19 175.18
- X1    1    13.615 671.97 175.97
<none>              658.36 176.38
+ X6    1     5.693 652.66 177.70
+ X2    1     1.385 656.97 178.21
- X5    1   201.768 860.12 195.23
- X4    1   220.994 879.35 196.95

Step:  AIC=175.18
Y ~ X1 + X4 + X5

       Df Sum of Sq    RSS    AIC
<none>              665.19 175.18
+ X2    1     7.029 658.16 176.35
+ X3    1     6.833 658.36 176.38
+ X6    1     1.972 663.22 176.95
- X1    1    37.595 702.78 177.47
- X5    1   198.891 864.08 193.59
- X4    1   282.896 948.08 200.82
```

Τα σύμβολο + (–) που εμφανίζεται μπροστά από τις μεταβλητές $X_1,\ X_2, ...,\ X_6$ δηλώνουν ότι η συγκεκριμένη μεταβλητή προστίθεται (αφαιρείται) από το εκάστοτε μοντέλο για να προκύψει η τιμή του AIC. Σύμφωνα με τα παραπάνω το καλύτερο μοντέλο για την περιγραφή των δεδομένων μας είναι το

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_5 + \varepsilon\ .$$

Στο ίδιο αποτέλεσμα καταλήγουμε εκτελώντας τις εντολές

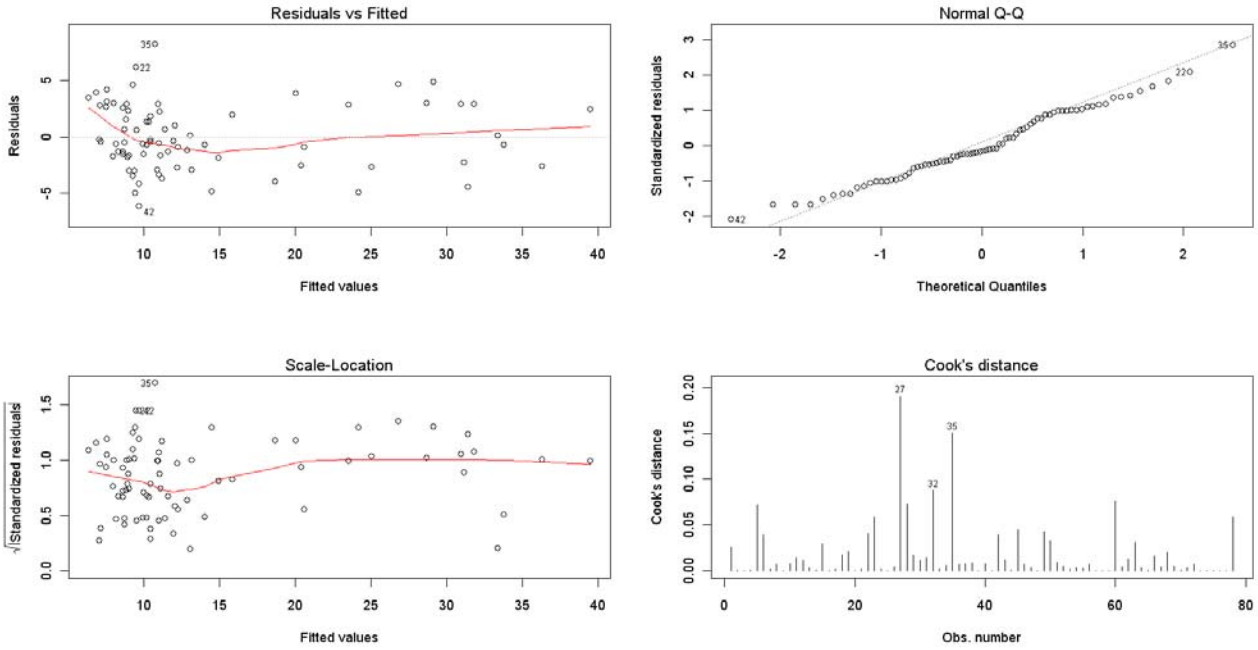
```
> mod2 <- stepAIC(m, direction="backward",k=2)
> mod3 <- stepAIC(lm(Y ~ 1), direction="forward",
+ scope=(~X6+X5+X4+X3+X2+X1),k=2)
```

Ωστόσο, αλλάζοντας την τιμή του k από 2 σε log(n) προκύπτει ότι το καταλληλότερο μοντέλο είναι το

$$Y = \beta_0 + \beta_1 x_4 + \beta_2 x_5 + \varepsilon$$

Με τη συνάρτηση `plot` λαμβάνουμε γραφικές παραστάσεις σχετικές με το μοντέλο μας (η ερμηνεία τους είναι ανάλογη με αυτή της απλής γραμμικής παλινδρόμησης).

```
> par(mfrow=c(2,2))
> plot(m, which=(1:4))
```



6.8 Σύνοψη εντολών Κεφαλαίου 6

`aov, ad.test, add1, anova`

`binom.test`

`chisq.test, confint, cor.test, cvm.test`

`drop1`

`fisher.test`

`kruskal.test, ks.test`

`levene.test, lillie.test, lm`

`pearson.test, predict, prop.test`

`qqline, qqnorm, qqplot`

`runs.test`

`sf.test, shapiro.test, SIGN.test, stepAIC`

`TukeyHSD, t.test`

`var.test, vif`

`wilcox.test`

ΚΕΦΑΛΑΙΟ 7

Προγραμματισμός στο R

7.1 Βασικά στοιχεία προγραμματισμού: Ομαδοποίηση, επανάληψη, και δεσμευμένη εκτέλεση εντολών

Η βασική δομή δεσμευμένης εκτέλεσης εντολών του R είναι η δομή `if` που έχει την ακόλουθη σύνταξη

```
if (λογική_έκφραση) {
    έκφραση_1
    ...
}
```

Η `λογική_έκφραση` στη δομή `if` πρέπει να επιστρέφει μια λογική τιμή, αληθής (`TRUE`) ή ψευδής (`FALSE`). Αν η λογική τιμή είναι αληθής (`TRUE`), τότε εκτελείται ότι υπάρχει εντός των άγκιστρων { ... }. Πολλές εκφράσεις που περικλείονται με άγκιστρα θεωρούνται ως μια έκφραση. Εάν υπάρχει μόνο μια έκφραση εντός των άγκιστρων τότε δεν είναι απαραίτητη η χρήση τους. Μια φυσική επέκταση της παραπάνω δομής `if` περιλαμβάνει και ένα `else` μέρος. Η σύνταξη είναι η ακόλουθη

```
if (λογική_έκφραση) {
    έκφραση_1
    ...
} else {
    έκφραση_2
    ...
}
```

Ενδιαφέρον έχει και η ακόλουθη δομή

```
if (λογική_έκφραση_1) {
    έκφραση_1
    ...
} else if (λογική_έκφραση_2) {
    έκφραση_2
    ...
} else {
    έκφραση_3
    ...
}
```

Για παράδειγμα

```
> x <- rnorm(1, 10, 1)
> a <- 10.5
> b <- pnorm(a, 10, 1)
> if (x<a) {
+ cat("The value", x, "is less than the", b, "percentile of N(10,1)", "\n")
+ } else {
+ cat("The value", x, "is greater than the", b, "percentile of N(10,1)", "\n")
+ }
The value 10.64724 is greater than the 0.6914625 percentile of N(10,1)
```

Αντί της δομής που βασίζεται στο `if`, χρησιμοποιείται αρκετά συχνά στην πράξη η δομή που βασίζεται στο `ifelse`, που έχει την ακόλουθη σύνταξη

```
ifelse (λογική έκφραση, A, B)
```

Η παραπάνω δομή επιστρέφει το `A` αν η `λογική_έκφραση` είναι αληθής, ειδάλλως επιστρέφει το `B`. Για παράδειγμα

```
> y <- -5:5
> z <- ifelse (y<0, -1, 1)
> z
 [1] -1 -1 -1 -1 -1  1  1  1  1  1  1
```

Για επαναληπτική εκτέλεση εντολών στο R χρησιμοποιείται η δομή `for` που έχει την ακόλουθη σύνταξη

```
for (x in vector) {
    έκφραση_1
    ...
}
```

Το `x` είναι η μεταβλητή του βρόγχου (loop), το `vector` είναι ένα διάνυσμα (ή μια έκφραση που ορίζει ένα διάνυσμα), και ότι υπάρχει μεταξύ των άγκιστρων { ... } εκτελείται για κάθε ένα στοιχείο του `vector`. Ακολούθως δίνονται τρία παραδείγματα:

```
> # Celsius to Fahrenheit
> for (celsius in 30:35) print(c(celsius, 9/5*celsius+32))
[1] 30   86
[1] 31.0 87.8
[1] 32.0 89.6
[1] 33.0 91.4
[1] 34.0 93.2
[1] 35   95
```

```
> # Άθροισμα συνιστωσών ενός διανύσματος
> vec <- seq(5,30,5)
> sum_x <- 0
> for (x in vec) {
+       sum_x <- sum_x+x
+  }
```

```
> cat("The vector is (",vec,")", "\n")
The vector is ( 5 10 15 20 25 30 )
> cat("The sum of the elements of the vector is", sum_x, "\n")
The sum of the elements of the vector is 105
```

```
> y <- -5:5
> for(i in 1:length(y)) {
+    if (y[i]<0) y[i]<- 0
+ }
> y
 [1] 0 0 0 0 0 0 1 2 3 4 5
> # Εναλλακτικός τρόπος
> y <- -5:5
> y[y<0] <- 0
> y
 [1] 0 0 0 0 0 0 1 2 3 4 5
```

Για τη δημιουργία βρόγχων χρήσιμη είναι και η δήλωση `while` που έχει την ακόλουθη σύνταξη

```
while (λογική_έκφραση) {
      έκφραση_1
      ...
}
```

Όταν εκτελείται μια εντολή `while` υπολογίζεται αρχικά η `λογική_έκφραση`. Εάν είναι αληθής, τότε εκτελείται η ομάδα των εντολών που περιέχονται μεταξύ των άγκιστρων { ... }. Ο έλεγχος μετά επιστρέφει στην αρχή της εντολής `while` και αν η `λογική_έκφραση` είναι και πάλι αληθής, η ομάδα των εντολών που περιέχονται μεταξύ των άγκιστρων εκτελείται ξανά, κοκ. Αν η `λογική_έκφραση` είναι ψευδής τότε σταματά να εκτελείται ο βρόγχος.

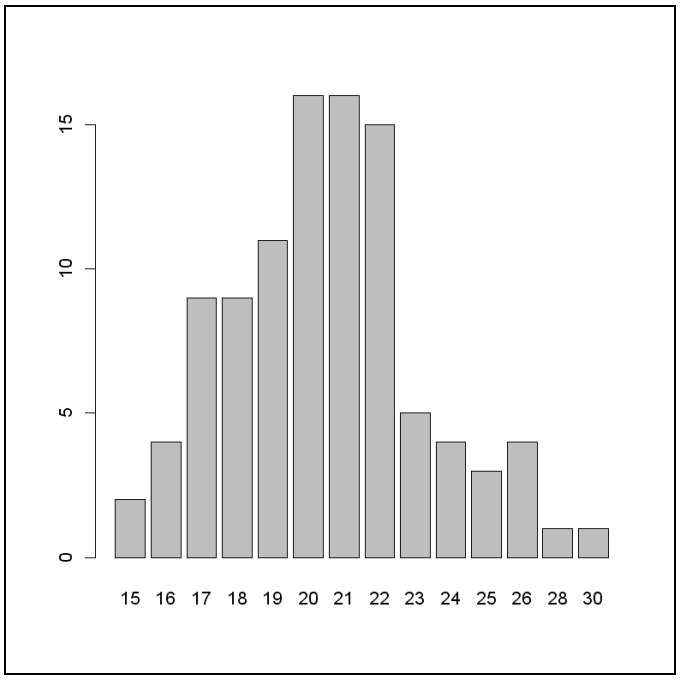
Σε μια δομή επαναληπτικής εκτέλεσης εντολών μπορεί να χρησιμοποιηθεί και η δήλωση `repeat` που συντάσσεται ως ακολούθως

```
repeat {
      έκφραση_1
      ...
}
```

Η δήλωση `repeat` είναι μια απλούστερη εκδοχή της δήλωσης `while`. Ανάλογα χειριζόμαστε τη δήλωση `break` που τερματίζει τον τρέχοντα βρόγχο, και τη δήλωση `next` που τερματίζει τον τρέχοντα βρόγχο και αρχίζει την επόμενη επανάληψη.

Στο ακόλουθο παράδειγμα δίνουμε ένα ραβδόγραμμα για τον ελάχιστο αριθμό των τυχαίων μεταβλητών που πρέπει να αθροίσουμε έτσι ώστε το άθροισμα να υπερβεί ένα αριθμό *y*, στην περίπτωση που κάθε μια τυχαία μεταβλητή ακολουθεί την ομοιόμορφη κατανομή στο διάστημα (0,1). Η σχετική διαδικασία επαναλαμβάνεται 100 φορές.

```
> y<-10
> x<-1:100
> for (i in 1:100) {
+       s<-0
+       n<-0
+       while(s<y){
+       s<-s+runif(1,0,1); n<-n+1
+ }
+       x[i]<-n
+ }
> barplot(table(x))
```



Για την καλύτερη κατανόηση των δομών που παρουσιάστηκαν στην παρούσα παράγραφο δίνουμε τα ακόλουθα απλά παραδείγματα:

```
> for (i in 1:100) {
+     print(c(i,i^2))
+     if (i^2>=50)
+     break
+ }
[1] 1 1
[1] 2 4
[1] 3 9
[1]  4 16
[1]  5 25
[1]  6 36
[1]  7 49
[1]  8 64
> # Εναλλακτικός τρόπος με στοίχιση των αποτελεσμάτων (συνάρτηση format)

> for (i in 1:100) {
+       sq <- i^2
+       cat(format(i, width=4),
+           format(sq, width=6),
+           "\n", sep="")
```

```
+      if (i^2>=50)
+      break
+  }
   1     1
   2     4
   3     9
   4    16
   5    25
   6    36
   7    49
   8    64
```

```
> x<-2; i<-0
> while(2*x < 1000) {
+       x<-x*2
+       i<-i+1
+       cat("x=",x,"  i=",i,"\n")
+ }
x= 4   i= 1
x= 8   i= 2
x= 16   i= 3
x= 32   i= 4
x= 64   i= 5
x= 128   i= 6
x= 256   i= 7
x= 512   i= 8
```

```
> x<-2; i<-0
> repeat{
+    i<-i+1
+    x<-2*x
+    cat("x=",x,"  i=",i,"\n")
+    if(2*x > 1000) break
+ }
x= 4   i= 1
x= 8   i= 2
x= 16   i= 3
x= 32   i= 4
x= 64   i= 5
x= 128   i= 6
x= 256   i= 7
x= 512   i= 8
```

Ορισμένες φορές προτιμούμε να γράφουμε μια λίστα εντολών σε ένα script αρχείο ή σε ένα πρόγραμμα οι οποίες θέλουμε να εκτελεστούν η μία κατόπιν της άλλης ανακαλώντας το συγκεκριμένο αρχείο/πρόγραμμα με ειδική εντολή. Αυτό γίνεται μέσω της συνάρτησης `source`. Αρχικά θα πρέπει να σώσουμε τις εντολές που θέλουμε να εκτελεστούν σε ένα αρχείο τύπου R (*.R). Επίσης, είναι αρκετά σύνηθες να γράφουμε εντολές για την εκτέλεση ενός υπολογισμού που να απαιτεί να εισάγουμε τις τιμές κάποιων παραμέτρων. Για παράδειγμα, η λύση της δευτεροβάθμιας εξίσωσης

$ax^2 + bx + c = 0$ εξαρτάται από τους συντελεστές των όρων x^2, x^1 και x^0 της εξίσωσης. Για να εισάγουμε τις τιμές των παραμέτρων με το πληκτρολόγιο χρησιμοποιούμε την συνάρτηση `readline` (ανάλογης φύσης εντολή ήταν και η συνάρτηση `scan` που συναντήσαμε στα διανύσματα στο Κεφάλαιο 2). Για να διαφωτίσουμε τη χρήση των συναρτήσεων `source` και `readline` θα κατασκευάσουμε ένα πρόγραμμα για τη λύση μιας δευτεροβάθμιας εξίσωσης. Οι κατάλληλες εντολές για τη λύση της δευτεροβάθμιας εξίσωσης περιέχονται στο αρχείο quad.R το οποίο πρέπει να βρίσκεται στον κατάλογο εργασίας μας.

```
# Αρχείο quad.R
# Roots of the quadratic equation a*x^2 + b*x + c = 0
cat("Roots of the equation a*x^2 + b*x + c = 0\n")

# Input
a <- as.numeric(readline("a = "))
b <- as.numeric(readline("b = "))
c <- as.numeric(readline("c = "))

# Discriminant
discrim <- b^2-4*a*c

# Roots
if (discrim > 0) {
    roots <- (-b + c(1,-1)*sqrt(b^2 - 4*a*c))/(2*a)
} else {
    if (discrim == 0) {
        roots <- -b/(2*a)
    } else {
        roots <- c()
    }
}

# Output
if (length(roots) == 0) {
    cat("The are no real roots\n")
} else if (length(roots) == 1) {
    cat("There is a double root: r =", roots, "\n")
} else {
    cat("There are two roots: r1 =", roots[1], " and  r2 =", roots[2], "\n")
}
```

Στη συνέχεια εκτελούμε την εντολή `source("quad.r")` όπου θα μας ζητηθούν οι τιμές των παραμέτρων *a*, *b* και *c* τις οποίες θα εισάγουμε με το πληκτρολόγιο (πατάμε το Enter μετά από κάθε εισαγωγή).

```
> source("quad.r")
Roots of the equation a*x^2 + b*x + c = 0
a = 2
b = -8
c = 1
There are two roots: r1 = 3.870829  and  r2 = 0.1291713
```

Στην επόμενη παράγραφο θα αντιληφθούμε ότι μια πιο κομψή παραλλαγή της παραπάνω διαδικασίας μπορεί να δοθεί ορίζοντας την κατάλληλη συνάρτηση (`function`).

Πιο σύνθετα παραδείγματα επίδειξης εντολών επανάληψης και δεσμευμένης εκτέλεσης θα δοθούν σε επόμενες παραγράφους.

7.2 Προγραμματισμός με συναρτήσεις

Το R παρέχει δυνατότητα ορισμού συναρτήσεων που έχουν τη μορφή

```
name <- function(όρισμα_1, όρισμα_2, ...) {
     έκφραση_1
     έκφραση_2
     ...
     return(αποτέλεσμα)
}
```

όπου name είναι το όνομα της συνάρτησης και ότι υπάρχει μέσα στα άγκιστρα { ... } επιστρέφει μια μόνο τιμή που είναι η τιμή της συνάρτησης που δηλώνεται με το return(αποτέλεσμα). Η δήλωση return(αποτέλεσμα) μπορεί να εμφανίζεται περισσότερες από μία φορές και τότε η τιμή της συνάρτησης ορίζεται από την δήλωση return(αποτέλεσμα) που θα εκτελεστεί πρώτη. Εάν δεν υπάρχει η δήλωση return(αποτέλεσμα) τότε η τιμή της συνάρτησης είναι η τιμή που προκύπτει από την εκτέλεση των εντολών που υπάρχουν μέσα στα άγκιστρα { ... }. Η συνάρτηση καλείται και υπολογίζεται με την εντολή

```
> name(όρισμα_1, όρισμα_2, ...)
```

Στο ακόλουθο παράδειγμα δίνεται η συνάρτηση max2 που βρίσκει το μεγαλύτερο αριθμό μεταξύ δύο αριθμών.

```
> max2 <- function(s,t) {
+  if(s>t) return(cat("The maximum between", s, "and", t, "is", s, "\n"))
+  if(t>s) return(cat("The maximum between", s, "and", t, "is", t, "\n"))
+  else print("The values are equal")
+ }
> max2(5,3)
The maximum between 5 and 3 is 5
> max2(7,15)
The maximum between 7 and 15 is 15
> max2(2,2)
The values are equal
> max2(e,7)
Error in max2(e, 7) : object 'e' not found"
```

Ένα άλλο απλό παράδειγμα είναι ο υπολογισμός του παραγοντικού ενός θετικού ακέραιου αριθμού. Δίνουμε τρεις διαφορετικές εκδοχές για τον υπολογισμό του με τις συναρτήσεις f1, f2 και f3.

```
> f1 <- function(x){
+       f<-1
+       if(x<2)return(1)
+       for(i in 2:x) {
+       f<- f*i
```

```
+ }
+       return(f)
+ }
> f1(5)
[1] 120
> sapply(0:5,f1)
[1]   1   1   2   6  24 120
```

```
> f2 <- function(x){
+       f<-1
+       t<-x
+       while(t>1){
+       f<-f*t
+       t<-t-1
+ }
+       return(f)
+ }
> f2(5)
[1] 120
> sapply(0:5,f2)
[1]   1   1   2   6  24 120
```

```
> f3 <- function(x){
+       f<-1
+       t<-x
+       repeat{
+       if (t<2) break
+       f<-f*t
+       t<-t-1
+ }
+       return(f)
+ }
> f3(5)
[1] 120
> sapply(0:5,f3)
[1]   1   1   2   6  24 120
```

Τα παραπάνω αποτελέσματα μπορούν να επιβεβαιωθούν άμεσα με τη χρήση της συνάρτησης `factorial` του R.

```
> sapply(0:5,factorial)
[1]   1   1   2   6  24 120
```

Σημειώνουμε επίσης ότι έγινε χρήση της συνάρτησης `sapply(x, FUN, ...)` (δείτε επίσης `apply`, `lapply`, `tapply` και `mapply`) η οποία εφαρμόζει τη συνάρτηση `FUN` σε κάθε στοιχείο του διανύσματος x και επιστρέφει αναλόγως ένα πίνακα ή ένα διάνυσμα.

Στη συνέχεια δίνουμε ένα τέταρτο τρόπο υπολογισμού του παραγοντικού ενός θετικού ακέραιου αριθμού με τεχνικές επαναληπτικού προγραμματισμού.

```
> f4 <- function(n) {
+     if (n == 1) {
+         return(1)
+     } else {
+         return(n*f4(n-1))
+     }
+ }
> f4(5)
[1] 120
```

Η ακόλουθη συνάρτηση ab διευκρινίζει ορισμένα σημεία σχετικά με τις τιμές των ορισμάτων (μεταβλητών) των συναρτήσεων, στην περίπτωση που έχουν προεπιλεγεί οι τιμές τους.

```
> ab <- function(x=6, y=4, z=2) {
+       w <- (x+y)/z
+       return(round(w, digits=3))
+ }
> ab()
[1] 5
> ab(10,20)
[1] 15
> ab(z=0.5)
[1] 20
> ab(8, ,3)
[1] 4
```

Ένα σημείο στο οποίο πρέπει να δοθεί προσοχή είναι στα ορίσματα και τις μεταβλητές οι οποίες εμφανίζονται μέσα στο "περιβάλλον" μιας συνάρτησης. Αυτά τα ορίσματα και οι μεταβλητές δεν είναι "ορατά" έξω από το περιβάλλον της συνάρτησης. Ωστόσο μεταβλητές που είναι ορισμένες εκτός του περιβάλλοντος της συνάρτησης είναι ορατές μέσα στο περιβάλλον της συνάρτησης. Τα παραπάνω αποσαφηνίζονται με το ακόλουθο παράδειγμα όπου η μεταβλητή *y* ορίζεται εντός του περιβάλλοντος της συνάρτησης και η μεταβλητή *z* εκτός αυτού.

```
> test <- function(x) {
+    y <- x*z
+    return(y)
+ }
> z <- 2
> test(4)
[1] 8
> x
Error: object 'x' not found
> y
Error: object 'y' not found
> z
[1] 2
> z <- 6
> test(4)
[1] 24
```

7.3 Εφαρμογές

Στην παρούσα παράγραφο δίνονται ορισμένες εφαρμογές που εμβαθύνουν στις τεχνικές των δύο προηγούμενων παραγράφων.

- **Αριθμοί Fibonacci**

Η ακολουθία των αριθμών 1, 1, 2, 3, 5, 8, 13, ... ονομάζεται ακολουθία Fibonacci. Κάθε όρος της ακολουθίας είναι ίσος με το άθροισμα των δύο προηγούμενων όρων. Η ακόλουθη συνάρτηση `fibonacci` υπολογίζει το n – οστό όρο της ακολουθίας.

```
> fibonacci <- function(n) {
+    a<-1
+    b<-0
+    while(n>0) {
+    swap<-a
+    a<-a+b
+    b<-swap
+    n<-n-1
+ }
+    b
+ }
> fibonacci(9)
[1] 34
> sapply(1:10,fibonacci)
 [1]  1  1  2  3  5  8 13 21 34 55
```

- **Ασθενής νόμος των μεγάλων αριθμών**

Αν $X_1, X_2, \ldots$ είναι μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με πεπερασμένη μέση τιμή μ και διακύμανση σ^2, τότε η ακολουθία των δειγματικών μέσων

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}, \quad n = 1, 2, \ldots$$

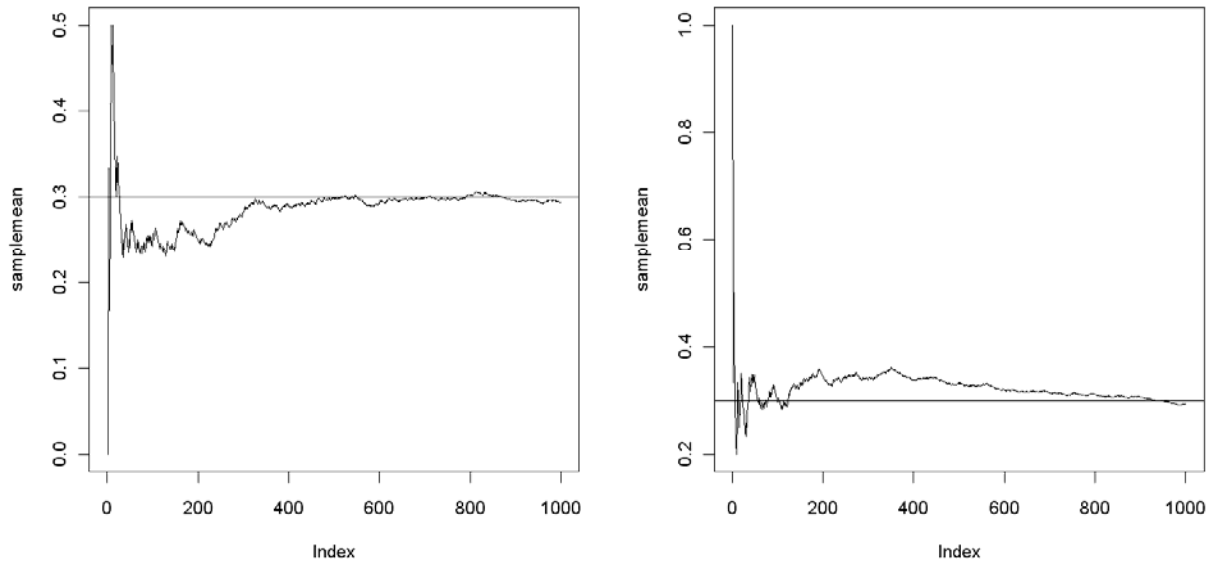
συγκλίνει κατά πιθανότητα στον κοινό μέσο μ των τυχαίων μεταβλητών $X_1, X_2, \ldots$, δηλαδή για κάθε θετικό αριθμό ε ισχύει ότι

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \varepsilon) = 0 .$$

Για παράδειγμα, αν $X_1, X_2, \ldots$ είναι μια ακολουθία ανεξάρτητων δοκιμών Bernoulli με πιθανότητα επιτυχίας p, τότε η ακολουθία των δειγματικών μέσων $\overline{X}_n$, $n = 1, 2, \ldots$, συγκλίνει κατά πιθανότητα προς τη σταθερά p αφού $E(X_1) = E(X_2) = \ldots = p$. Για να διαπιστώσουμε εμπειρικά το παραπάνω αποτέλεσμα θα δημιουργήσουμε δύο ακολουθίες που κάθε μια αποτελείται από 1000 ανεξάρτητες και ισόνομες δοκιμές Bernoulli με πιθανότητα επιτυχίας 0.3 (αντικείμενο `sequence`). Για τη δημιουργία κάθε ακολουθίας χρησιμοποιούμε 1000 τυχαίους αριθμούς από την ομοιόμορφη κατανομή στο διάστημα (0, 1). Στη συνέχεια κατασκευάζονται δύο index plot των δειγματικών μέσων

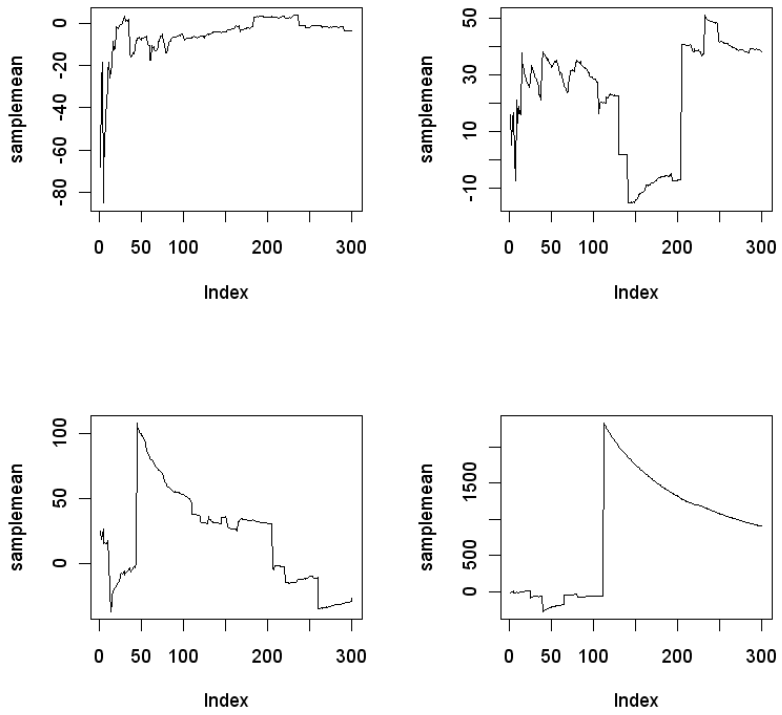
(αντικείμενο samplemean). Εναλλακτικά, θα μπορούσαμε άμεσα να σχηματίσουμε τις δύο ακολουθίες με την εντολή "> sequence <- rbinom(n,1,0.3)", αντί του πλάγιου τρόπου με χρήση ομοιόμορφης κατανομής.

```
> par(mfrow=c(1,2))
> n <- 1000
> for(i in 1:2) {
+     rs <- runif(n,0,1)
+     sequence <- as.numeric(rs>0.7)          # Εναλλακτικά rs<0.3
+     samplemean <- cumsum(sequence)/1:n      # «Πηλίκο» διανυσμάτων
+     plot(samplemean, type="l")
+     abline(0.3,0)
+ }
```



Ωστόσο στην περίπτωση της κατανομής Cauchy όπου η μέση τιμή δεν υπάρχει, περιμένουμε να μη συγκλίνει η ακολουθία των δειγματικών μέσων. Πράγματι

```
> par(mfrow=c(2,2))
> n <- 300
> for(i in 1:4) {
+     sequence <- rcauchy(n, 10, 20)
+     samplemean <- cumsum(sequence)/1:n
+     plot(samplemean, type="l")
+ }
```



- **Κεντρικό οριακό θεώρημα**

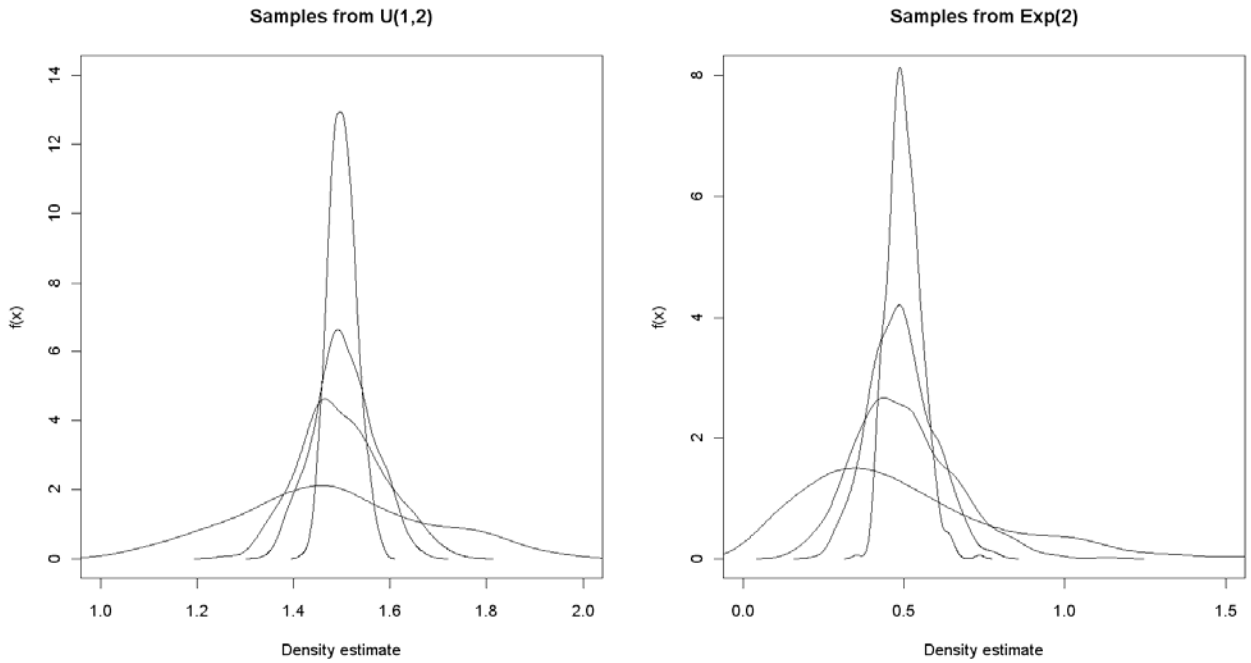
Αν $X_1, X_2, \ldots$ είναι μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με πεπερασμένη μέση τιμή μ και διακύμανση σ^2, τότε ο δειγματικό μέσος

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$\overline{X}_n$ ακολουθεί προσεγγιστικά την κανονική κατανομή $N(\mu, \sigma^2/n)$ για μεγάλες τιμές του n, σύμφωνα με μια απλοποιημένη διατύπωση του Κεντρικού Οριακού Θεωρήματος. Στο ακόλουθο παράδειγμα διαπιστώνουμε εμπειρικά το παραπάνω αποτέλεσμα βασιζόμενοι σε δείγματα από την ομοιόμορφης κατανομής στο διάστημα (1, 2) και την εκθετική κατανομή με παράμετρο $\lambda = 2$. Για κάθε κατανομή κατασκευάζεται διάνυσμα 500 δειγματικών μέσων (διάνυσμα `res`) το οποίο προκύπτει από 500 δείγματα μεγέθους $n = 2, 10, 25$ και 100.

```
> par(mfrow=c(1,2))
# Ομοιόμορφη κατανομή
> plot(0,0, type="n", xlim=c(1, 2), ylim=c(0, 14),
+ main="Samples from U(1,2)", xlab="Density estimate", ylab="f(x)")
> m <- 500; a <- 1; b <- 2;  res <- c()
> for(n in c(2, 10, 25, 100)) {
+     for (i in 1:m) res[i] <- mean(runif(n,a,b))
+     lines(density(res), lwd=1.5)
+ }
# Εκθετική κατανομή
> plot(0,0, type="n", xlim=c(0, 1.5), ylim=c(0, 8),
```

```
+ main="Samples from Exp(2)", xlab="Density estimate", ylab="f(x)")
> m <- 500; rate <- 2;  res <- c()
> for(n in c(2, 10, 25, 100)) {
+     for (i in 1:m) res[i] <- mean(rexp(n,rate))
+     lines(density(res), lwd=1.5)
+ }
```



- **Monte Carlo ολοκλήρωση**

Έστω X μια τυχαία μεταβλητή με κατανομή $U(a,b)$ και g μια (μετρήσιμη) συνάρτηση. Η μέση τιμή της τυχαίας μεταβλητής $g(X)$ δίνεται από τον τύπο

$$E[g(X)] = \int_a^b g(x) f(x) dx = \frac{1}{b-a} \int_a^b g(x) dx$$

οπότε

$$I = \int_a^b g(x) dx = (b-a) E[g(X)].$$

Σύμφωνα με τον ισχυρό νόμο των μεγάλων αριθμών, αν $X_1, X_2, \ldots$ είναι μια ακολουθία ανεξάρτητων τυχαίων μεταβλητών με κατανομή όπως της τυχαίας μεταβλητής X, τότε

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \underset{n \to \infty}{\rightarrow} E[g(X)]$$

με πιθανότητα 1. Συνεπώς μπορούμε να υπολογίσουμε προσεγγιστικά το ολοκλήρωμα I μέσω του προσεγγιστικού υπολογισμού της μέσης τιμής $E[g(X)]$. Ως παράδειγμα δίνεται ο προσεγγιστικός υπολογισμός του ολοκληρώματος

$$I = \int_0^2 e^{x^3} dx \, .$$

```
> a <- 0; b <- 2          # Όρια ολοκλήρωσης
> mci <- function(n) {
+     x <- runif(n,min=a,max=b)
+     y <- exp(x^3)
+     (b-a)*mean(y)
+ }
> cat("For n=10000 we have that I=", mci(10000),"\n")
For n=10000 we have that I= 278.8845
> cat("For n=1000000 we have that I=", mci(1000000),"\n")
For n=1000000 we have that I= 276.2181
```

Ο ακριβής υπολογισμός του I γίνεται με τη συνάρτηση `integrate` ως εξής:

```
> z <- function(x) {exp(x^3)}
> integrate(z, lower = 0, upper = 2)
276.8529 with absolute error < 3.2e-07
```

- **Γραφική απεικόνιση διαστημάτων εμπιστοσύνης του μέσου ενός κανονικού πληθυσμού**

Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα από κανονικό πληθυσμό με μέση τιμή μ και τυπική απόκλιση σ. Το διάστημα

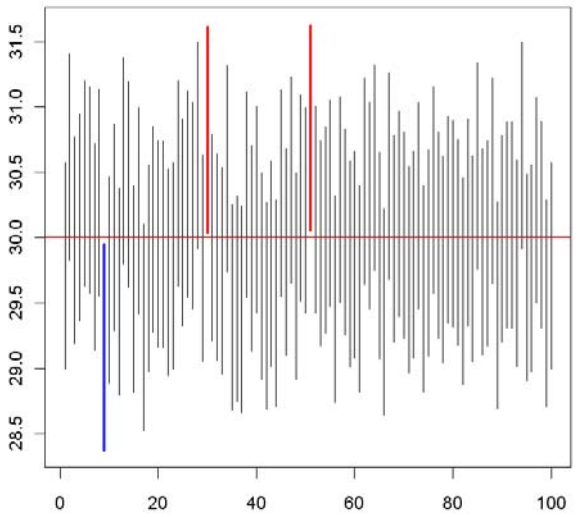
$$[L, U] = [\bar{x} - z_{a/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{a/2}\frac{\sigma}{\sqrt{n}}] = [\bar{x} - z_{(1-cl)/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{(1-cl)/2}\frac{\sigma}{\sqrt{n}}]$$

αποτελεί το $100(1-a)\% = 100cl\%$ διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού. Η ποσότητα $cl = 1-a$ ονομάζεται συντελεστής εμπιστοσύνης του διαστήματος. Αν ληφθεί ένας μεγάλος αριθμός τυχαίων δειγμάτων από τον πληθυσμό και για κάθε ένα δείγμα υπολογιστεί το $100(1-a)\%$ διάστημα εμπιστοσύνης, τότε περίπου το $100(1-a)\%$ από αυτά θα περιέχει το μέσο του πληθυσμού. Με την ακόλουθη συνάρτηση `CI.plot` μπορούμε να κάνουμε γραφική παράσταση *m* διαστημάτων εμπιστοσύνης που το καθένα θα απεικονίζεται ως ένα κάθετο ευθύγραμμο τμήμα με μήκος όσο το μήκος του διαστήματος. Επίσης καταγράφεται και ο συνολικός αριθμός των διαστημάτων που περιέχουν το μέσο.

```
CI.plot <- function (L, U) {
   plot(L, type = "n", ylim=c(min(L),max(U)), xlab = " ", ylab = " ")
   condition <- (L <= mu & U >= mu)
   segments((1:m)[L < mu & U > mu], L[L < mu & U > mu], (1:m)[L < mu & U > mu], U[L < mu & U > mu])
   segments((1:m)[L > mu], L[L > mu], (1:m)[L > mu], U[L > mu], col = "red", lwd = 3)
   segments((1:m)[U < mu], L[U < mu], (1:m)[U < mu], U[U < mu], col = "blue", lwd = 3)
   s <- sum(condition)
   abline(h = mu, col="brown", lwd=2)
   cat("Number of intervals that contain the true mean : ", s, "\n")
}
```

Με τις ακόλουθες εντολές ορίζονται 100 τυχαίοι δειγματικοί μέσοι (διάνυσμα a) από τον ίδιο κανονικό πληθυσμό και τα αντίστοιχα $100(1-a)\%$ διαστήματα εμπιστοσύνης $[L, U]$ (διανύσματα L και U).

```
> mu<-30    # Population mean
> std<-9    # Population standard deviation
> cl<-0.95  # Confidence level
> m<-100    # Number of samples
> n<-500    # Sample size
> z<-qnorm((1+cl)/2)
> a<-rep(0, times=m)
> L<-rep(0, times=m)
> U<-rep(0, times=m)
> i<-0
> while (i<m) {i<-i+1
+ a[i]<-mean(rnorm(n, mean = mu, sd = std))  # The i-th sample mean
+ L[i]<-a[i]-z*std*sqrt(1/n)  # The i-th lower limit
+ U[i]<-a[i]+z*std*sqrt(1/n)} # The i-th upper limit
> CI.plot(L,U)
Number of intervals that contain the true mean :  97
```



7.4 Σύνοψη εντολών Κεφαλαίου 7

break

else

factorial, for, format, function

if, ifelse, integrate

next

sapply, source

readline, repeat, return

while