Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

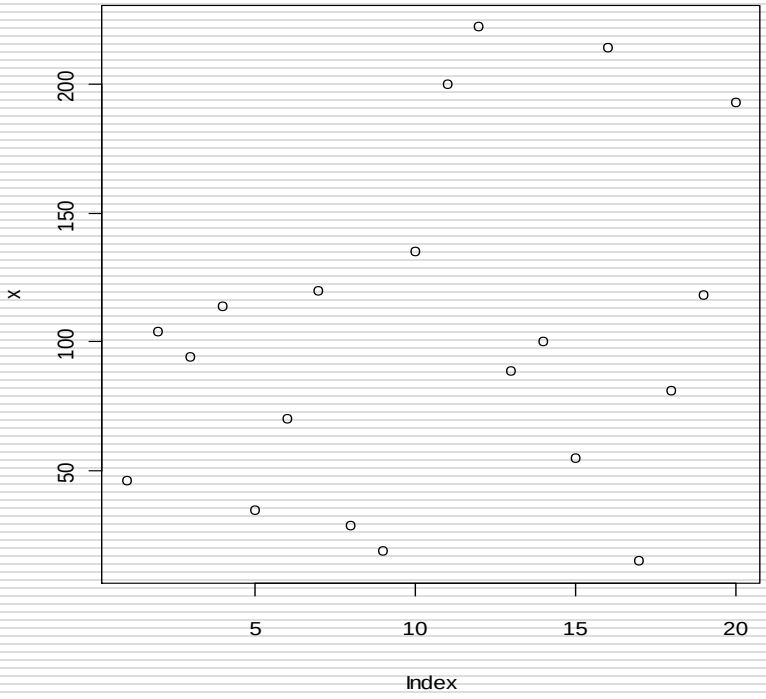
Δημήτρης Φουσκάκης,
Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

Απλά Διαγράμματα

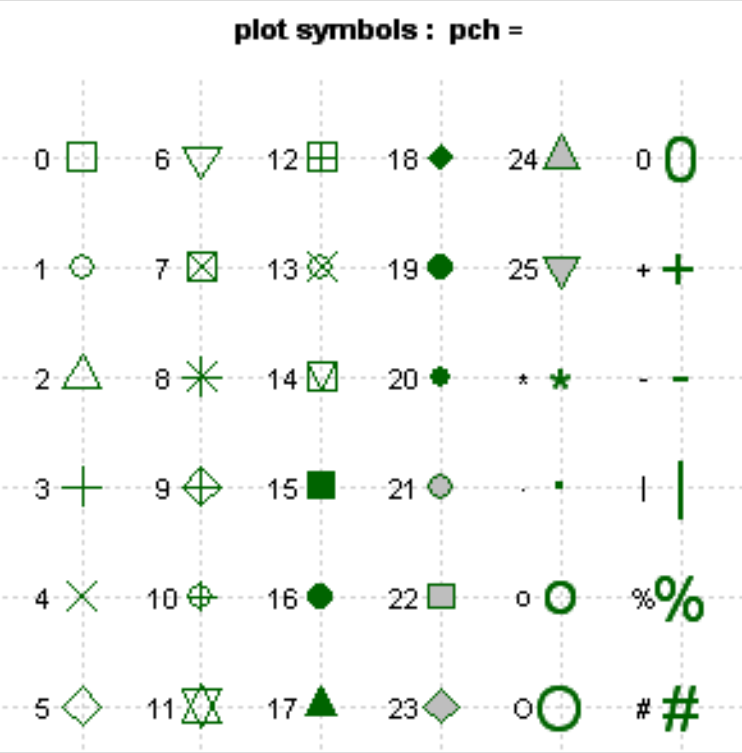
- Με την εντολή plot, μπορούμε να αναπαραστήσουμε γραφικά τις τιμές ενός διανύσματος. Για τα δεδομένα x του 1ου παραδείγματος, έχουμε

  >plot(x)



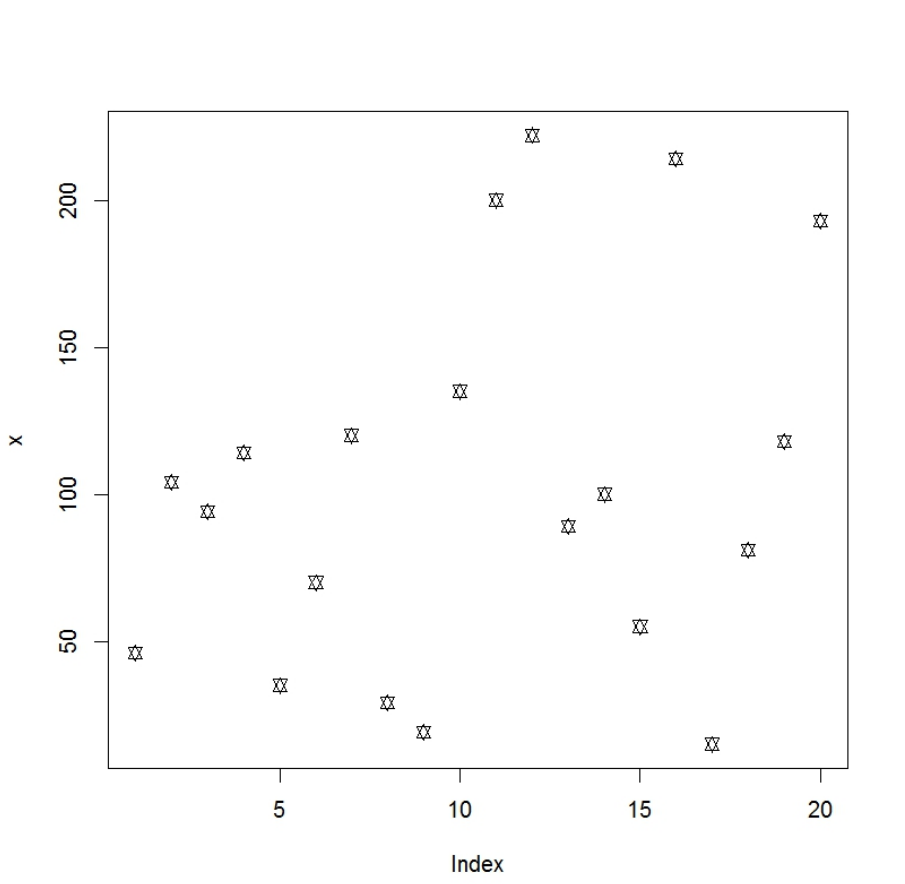
Απλά Διαγράμματα

- Μπορούμε να χρησιμοποιήσουμε διαφορετικά είδη συμβόλων με το όρισμα pch.



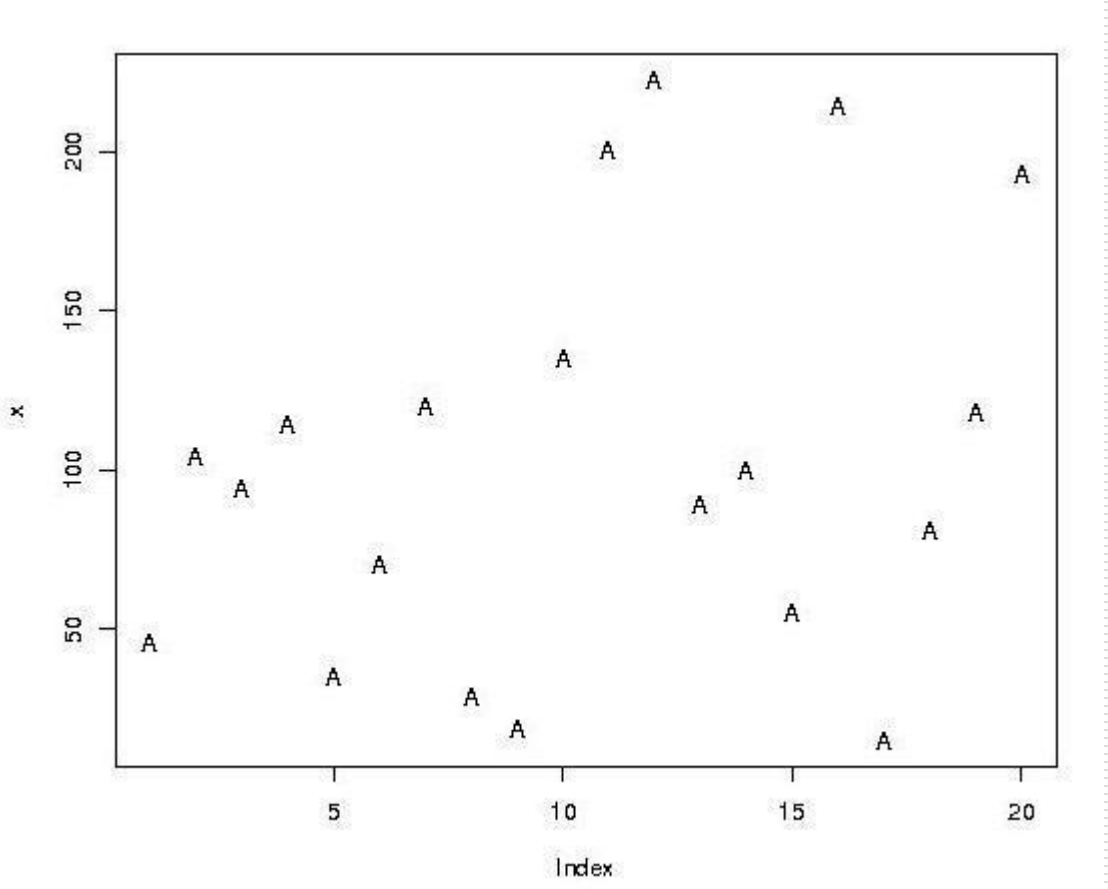
Απλά Διαγράμματα

plot(x, pch=11)



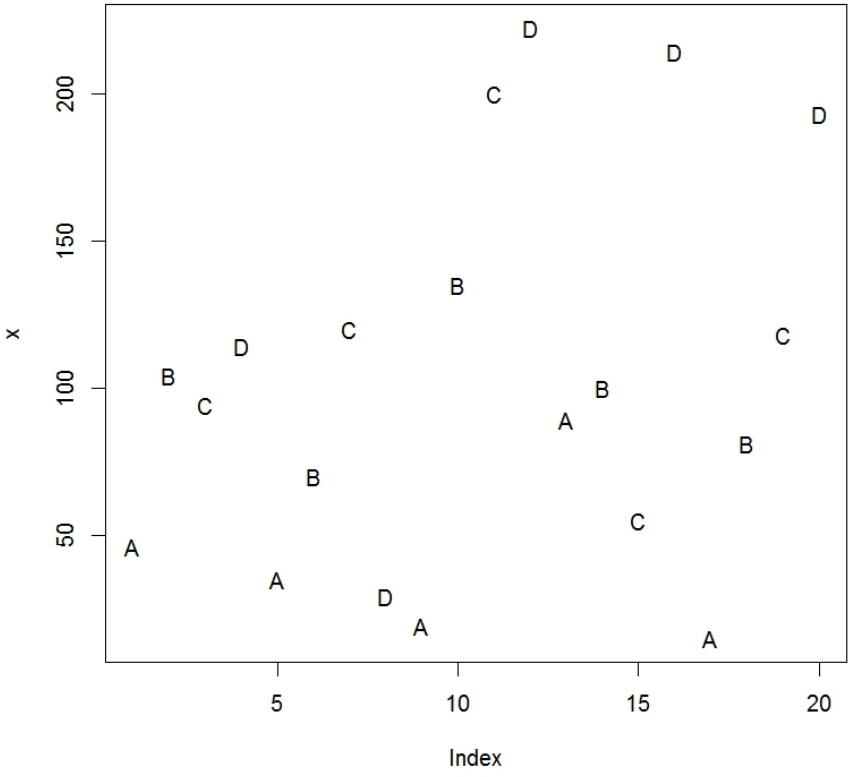
Απλά Διαγράμματα

plot(x,pch='A')



Απλά Διαγράμματα

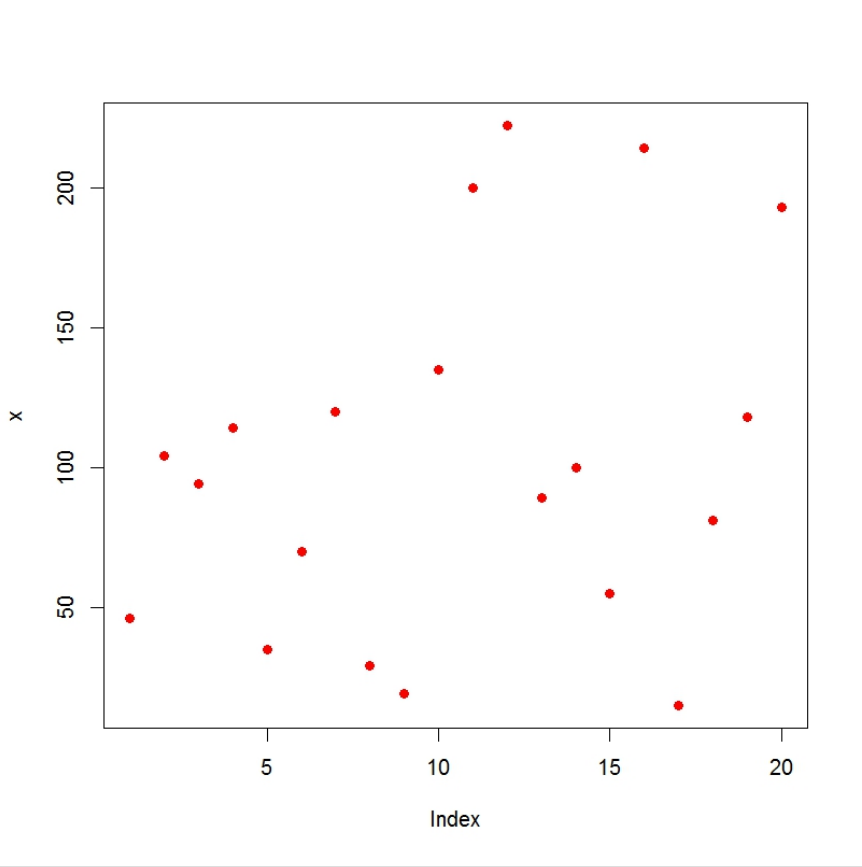
```
plot(x,pch=c('A','B'
,'C','D'))
```



Απλά Διαγράμματα

- Διαφορετικά χρώματα στα σύμβολα χρησιμοποιώντας το όρισμα col.

```
> plot(x,pch=16,col="red")
```



Απλά Διαγράμματα

- Οι διάφορες επιλογές χρωμάτων φαίνονται παρακάτω

```
> colors()
  [1] "white"             "aliceblue"         "antiquewhite"
  [4] "antiquewhite1"     "antiquewhite2"     "antiquewhite3"
  [7] "antiquewhite4"     "aquamarine"        "aquamarine1"
```

.....

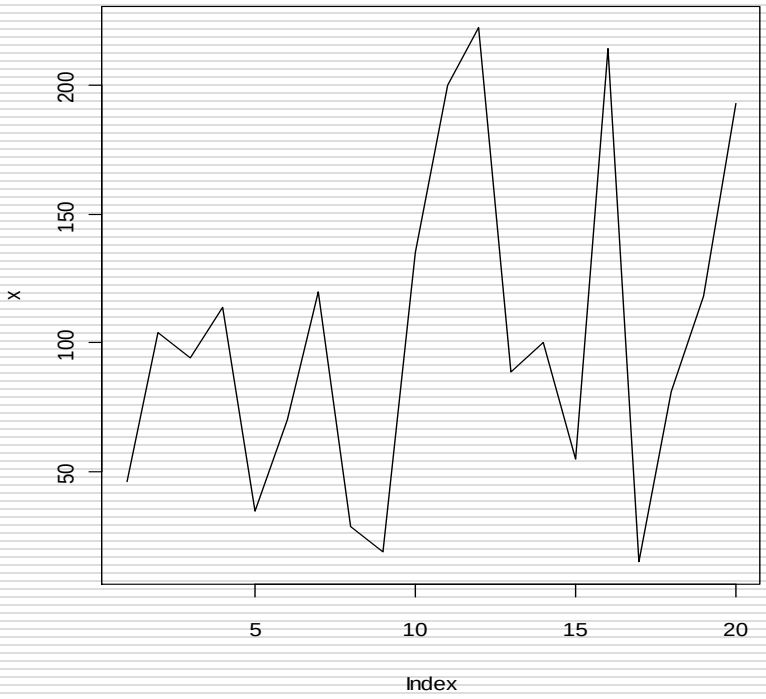
Αντιστοιχία των αριθμών 1 έως 8 σε χρώματα

```
> palette()
[1] "black"   "red"     "green3"  "blue"    "cyan"    "magenta"
"yellow"
[8] "gray"
```

Απλά Διαγράμματα

- Τα δεδομένα που αναπαριστάνουμε μπορούν να απεικονιστούν με διάφορους τρόπους με την βοήθεια του ορίσματος type. Π.χ.

```
> plot(x, type='l')
```

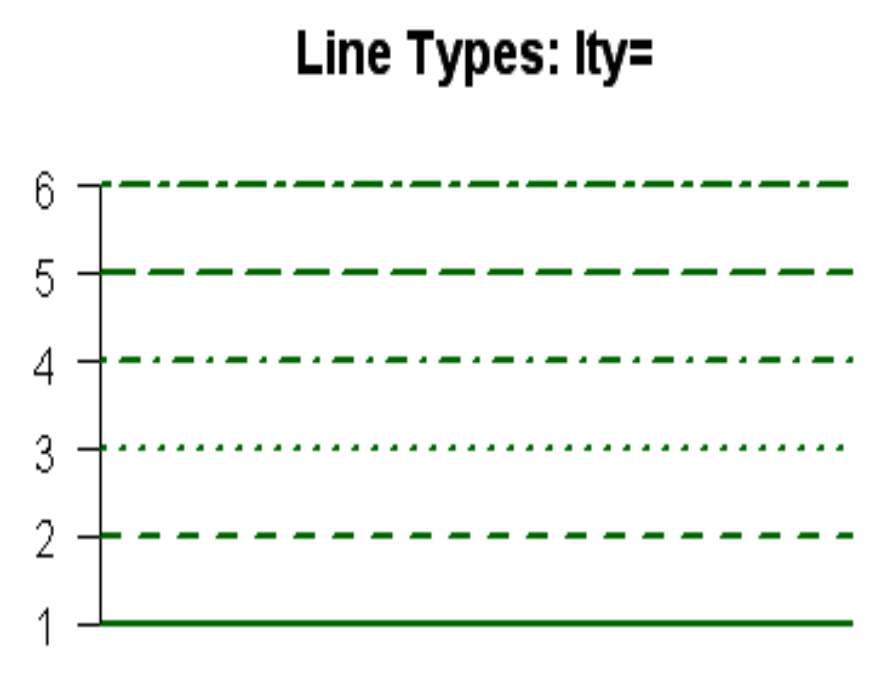


Απλά Διαγράμματα

Σύμβολο	Σημασία
p	Σημεία
l	Γραμμή
b	Γραμμή και σημεία
c	Γραμμή με κενό στα σημεία
o	Γραμμή και σημεία ενωμένα
h	Κάθετες γραμμές για κάθε σημείο
s	Με βήμα
n	Τίποτα

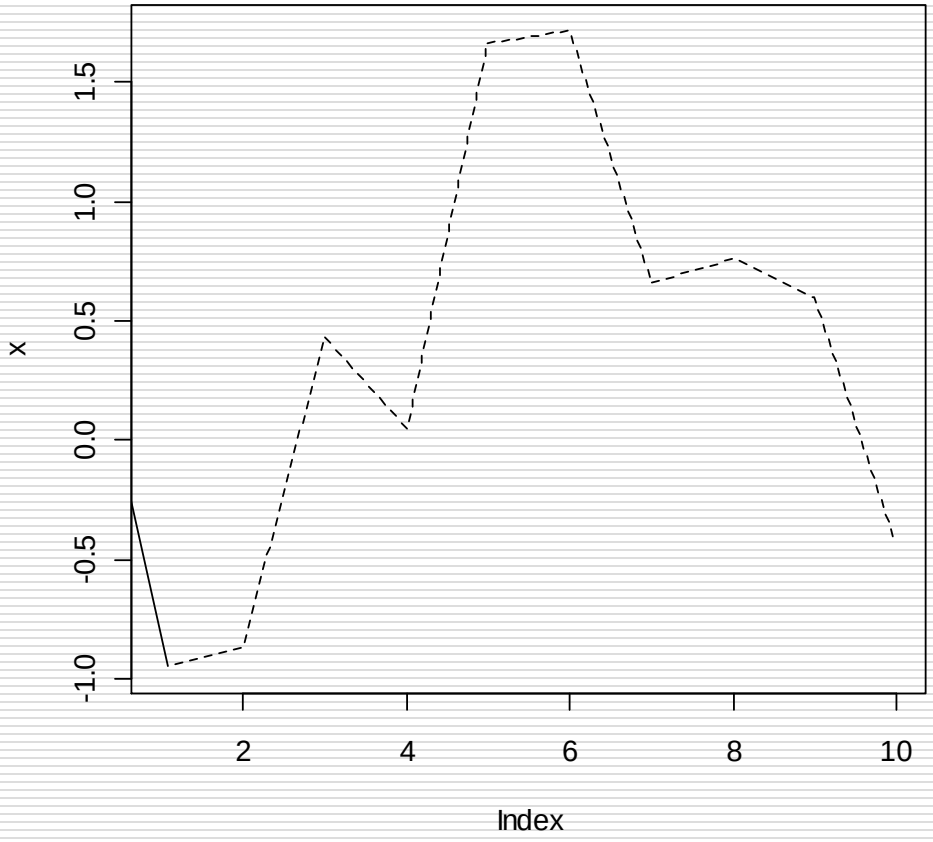
Απλά Διαγράμματα

- Μπορούμε να χρησιμοποιήσουμε διαφορετικά είδη γραμμών με το όρισμα lty.



Απλά Διαγράμματα

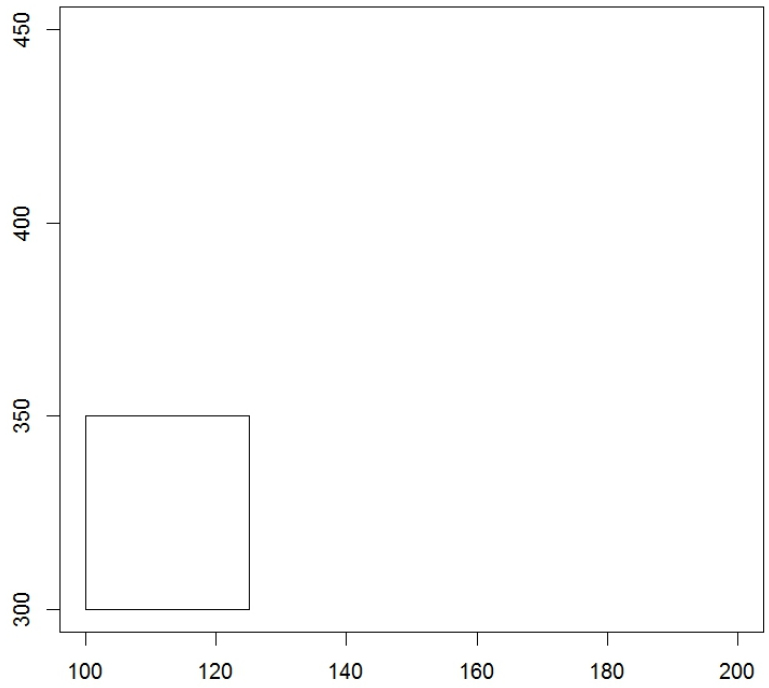
```
> plot(x, type='l', lty=2)
```



Απλά Διαγράμματα

- Μπορούμε να προσθέσουμε ένα ορθογώνιο παραλληλόγραμμο σε ένα διάγραμμα με την εντολή rect()

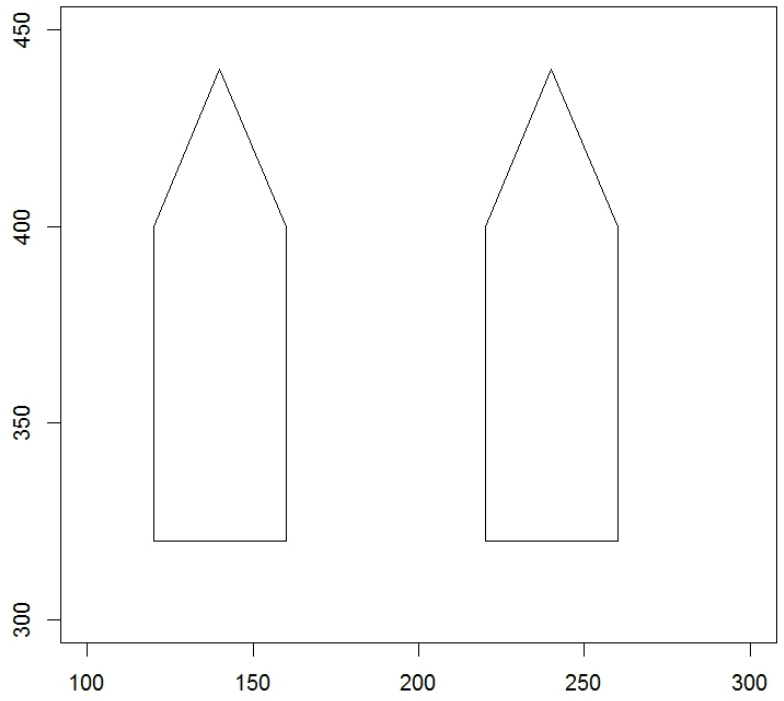
```
> plot(c(100,200), c(300,450), type="n", xlab="", ylab="")
> rect(100,300,125,350)
```



Απλά Διαγράμματα

- Με την εντολή polygon(x,y) μπορούμε να προσθέσουμε ένα ή περισσότερα πολύγωνα σε ένα διάγραμμα.

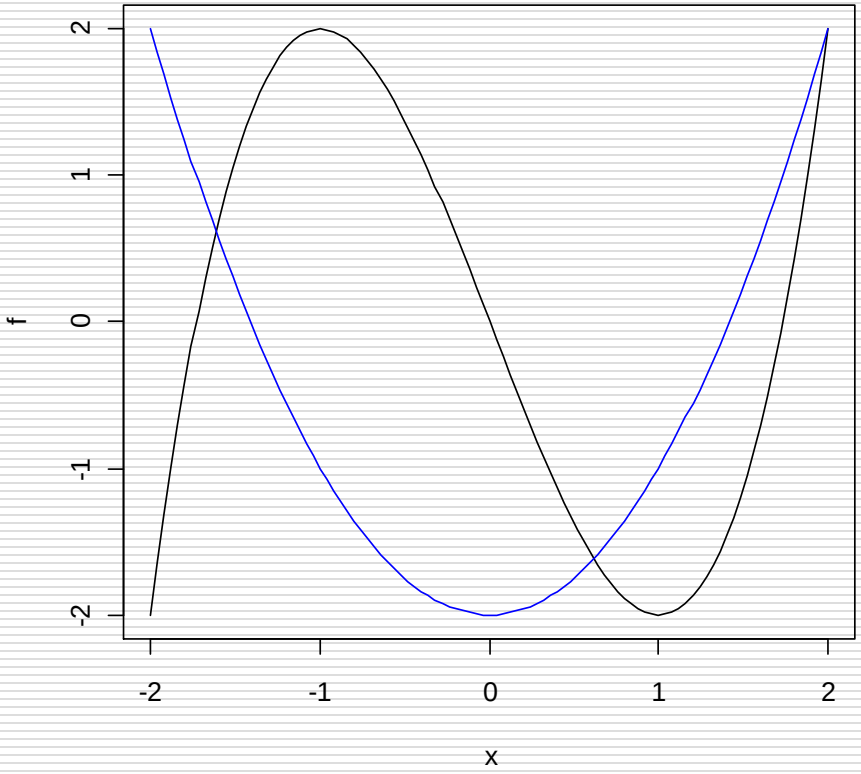
```
> plot(c(100,300), c(300,450),
    type="n", xlab="", ylab="")
> polygon(c(140,120,120,160,
160,NA,240,220,220,260,260),
    c(440,400,320,320,400,
    NA,440,400,320,320,400))
```



Απλά Διαγράμματα

- Μπορούμε να σχεδιάσουμε καμπύλες χρησιμοποιώντας την εντολή curve().

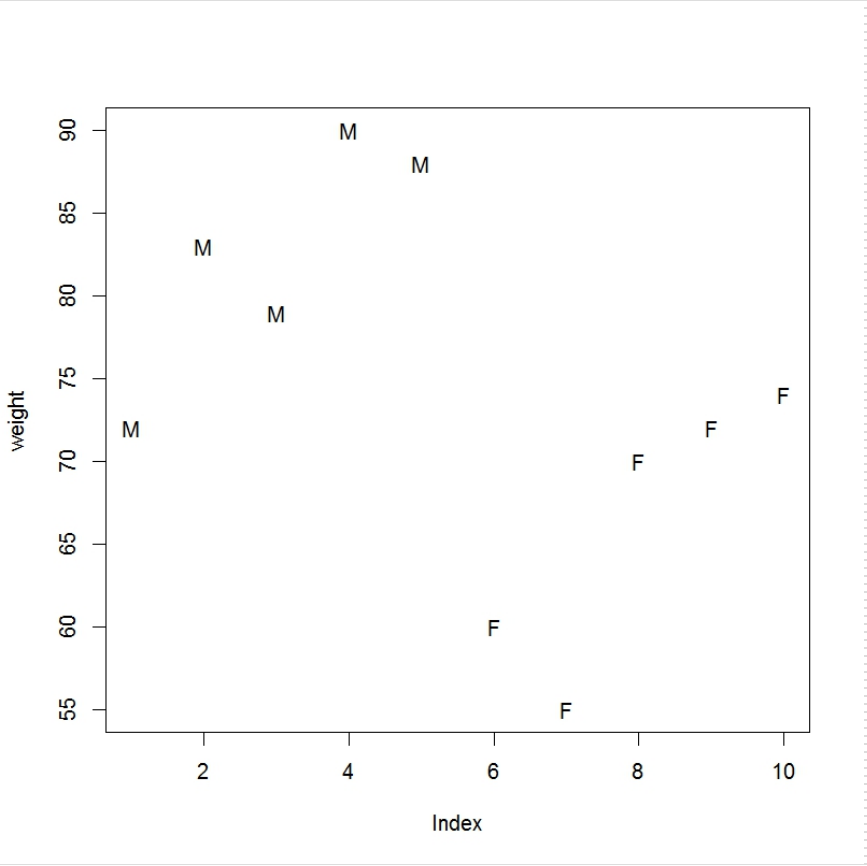
```
> plot(c(100,300),c(300,450),
    type="n",xlab="",ylab="")
> curve(x^3-3*x, -2, 2,
    ylab="f")
> curve(x^2-2, add=TRUE,
    col="blue")
```



Απλά Διαγράμματα

- Μπορούμε να εμφανίσουμε χαρακτήρες αντί για σημεία με την εντολή text()

```
> weight<-
c(72,83,79,90,88,60,55,70,72,
74)
> gender<-
rep(c("M","F"),each=5)
> gender
 [1] "M" "M" "M" "M" "M" "F"
"F" "F" "F" "F"
> plot(weight, type="n")
> text(weight, label=gender)
```

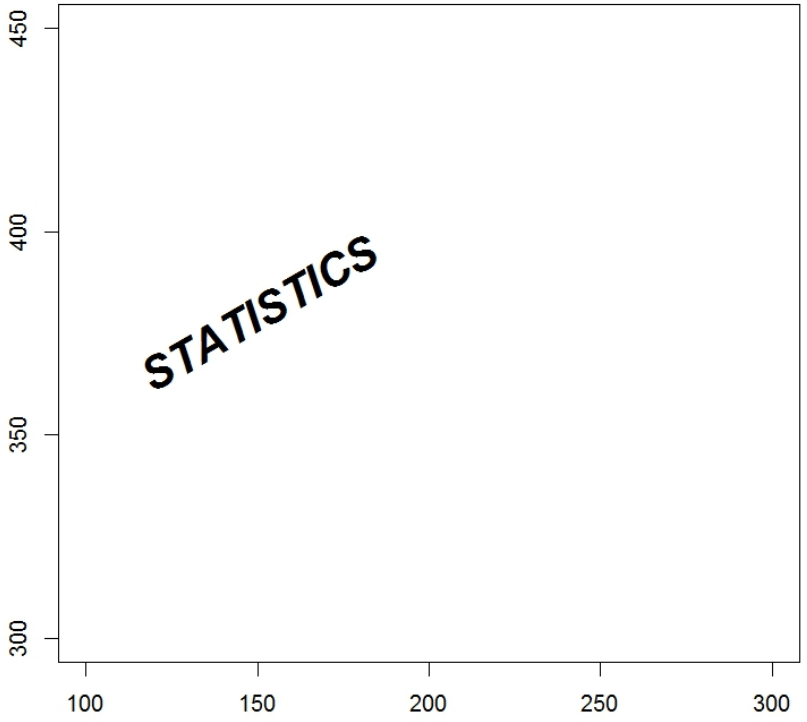


Απλά Διαγράμματα

- Περιστροφή χαρακτήρων με χρήση του ορίσματος srt

```
> plot(c(100, 300), c(300, 450),
    type= "n", xlab="", ylab="")
> text(150, 380, "STATISTICS",
    srt=30, cex=2, font=4)
```

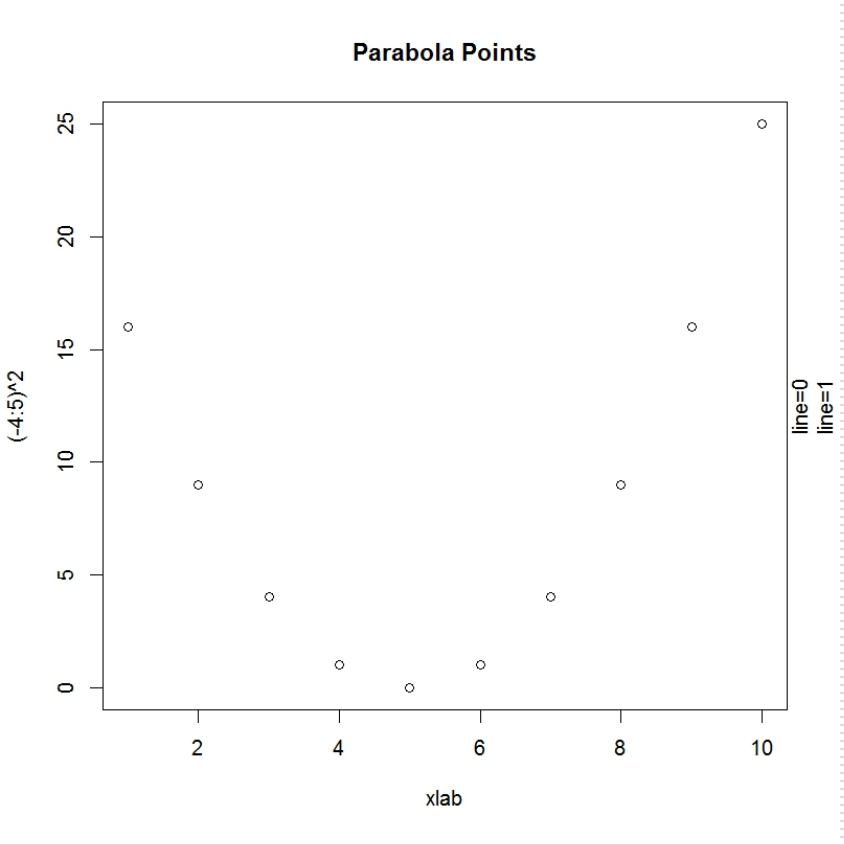
γωνία περιστροφής σε μοίρες



Απλά Διαγράμματα

- Με την εντολή mtext() μπορούμε να προσθέσουμε κείμενο σε οποιοδήποτε από τα τέσσερα περιθώρια του γραφικού παραθύρου.

```
> plot(1:10, (-4:5)^2, main="Parabola Points", xlab="xlab")
> mtext("line=0", side=4, line=0)
> mtext("line=1", side=4, line=1)
```

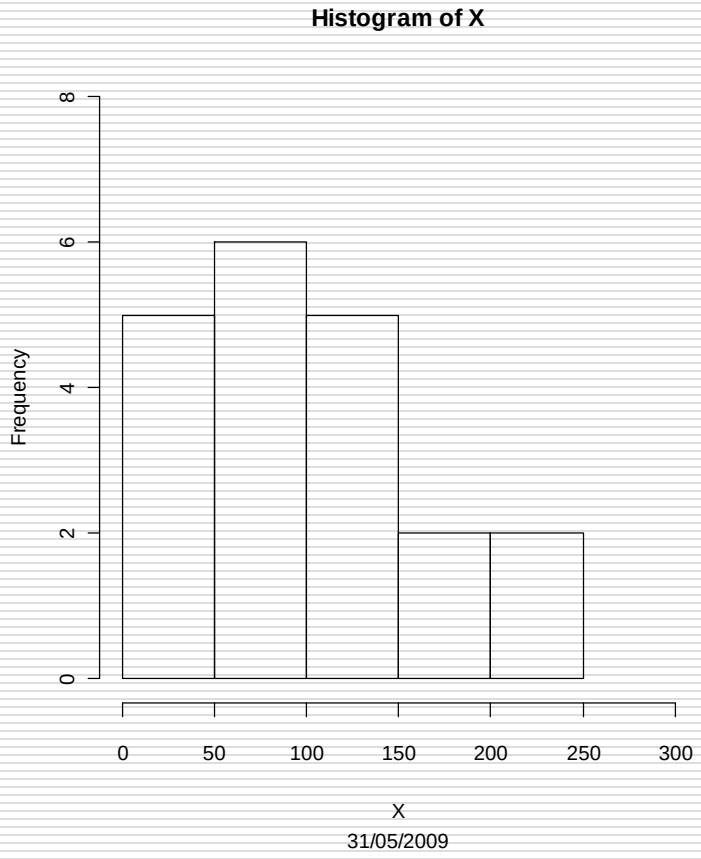


Περισσότερα στα Γραφήματα

- Σε κάθε εντολή δημιουργίας γραφημάτων μπορούν να δοθούν ορίσματα βελτίωσης της εικόνας τους.
  - Με την παράμετρο main δίνουμε τίτλο στο γράφημα.
  - Με την παράμετρο submain δίνουμε υπότιτλο στο γράφημα.
  - Με τις παραμέτρους xlab και ylab δίνουμε τίτλους στους άξονες.
  - Με την παράμετρο xlim και ylim δίνουμε επιθυμητό εύρος τιμών για τους άξονες.

Περισσότερα στα Γραφήματα

```
> hist(x,main="Histogram of X",
 sub="31/05/2009", xlab="X",
ylab="Frequency",
 ylim=c(0,8), xlim=c(0,300))
```



Περισσότερα στα Γραφήματα

- Όλες αυτές οι παράμετροι μπορούν να δοθούν σε ένα ήδη υπάρχον γράφημα με την βοήθεια της εντολής title.

```
> hist(x)
> title(main="Histogram of X", sub="31/05/2009", xlab="X",
    ylab="Frequency", ylim=c(0,8), xlim=c(0,300))
```

Μέγεθος των συμβόλων ή των χαρακτήρων

cex	**Δηλώνει το ποσοστό κατά το οποίο θα εμφανίζονται τα σύμβολα σε σχέση με το κανονικό (default), default=1 π.χ. 1.5 είναι 50% μεγαλύτερο από το κανονικό.**
cex.axis	Με ανάλογο τρόπο αλλάζουμε το μέγεθος των τιμών των αξόνων
cex.lab	Με ανάλογο τρόπο αλλάζουμε το μέγεθος των τιμών των τίτλων των αξόνων
cex.main	Με ανάλογο τρόπο αλλάζουμε το μέγεθος του τίτλου του διαγράμματος
cex.sub	Με ανάλογο τρόπο αλλάζουμε το μέγεθος του υπότιτλου του διαγράμματος

Χρώμα των συμβόλων ή των χαρακτήρων

col	**Δηλώνει το χρώμα στα σύμβολα ή στους χαρακτήρες ενός διαγράμματος.**
col.axis	Δηλώνει το χρώμα στις τιμές των αξόνων.
col.lab	Δηλώνει το χρώμα στους τίτλους των αξόνων.
col.main	Δηλώνει το χρώμα στον τίτλο του διαγράμματος.
col.sub	Δηλώνει το χρώμα στον υπότιτλο του διαγράμματος.
fg	Δηλώνει το χρώμα που χρησιμοποιείται στους άξονες και στο περίγραμμα του διαγράμματος (plot foreground color)
bg	Δηλώνει το χρώμα που χρησιμοποιείται για το φόντο του διαγράμματος (plot background color)

Στυλ της γραμματοσειράς για τα σύμβολα ή τους χαρακτήρες

font	**Δηλώνει το στυλ της γραμματοσειράς στα σύμβολα ή στους χαρακτήρες ενός διαγράμματος.**
font.axis	Δηλώνει το στυλ της γραμματοσειράς για τις τιμές των αξόνων.
font.lab	Δηλώνει το στυλ της γραμματοσειράς για τους τίτλους των αξόνων.
font.main	Δηλώνει το στυλ της γραμματοσειράς για τον τίτλο του διαγράμματος.
font.sub	Δηλώνει το στυλ της γραμματοσειράς για τον υπότιτλο του διαγράμματος.
family	Δηλώνει την οικογένεια της γραμματοσειράς που χρησιμοποιείται σε ένα διάγραμμα. Κάποιες συνηθισμένες τιμές είναι “serif”, “sans”, “mono”, “symbol”.

Μέγεθος περιθωρίου και διαγράμματος

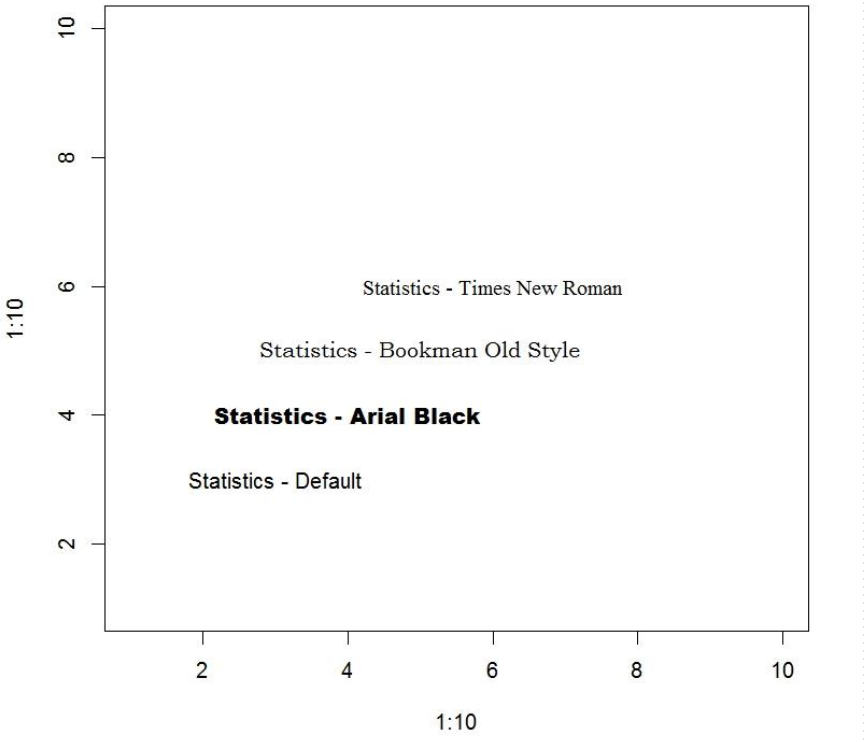
mar	**Παίρνει ως τιμές ένα αριθμητικό διάνυσμα, στο οποίο δηλώνεται το μέγεθος του περιθωρίου προσδιορισμένο με νοητές γραμμές. c(bottom , left, top, right) Π.χ. Bottom: ο αριθμός των νοητών γραμμών που θα απεικονίζονται στο κάτω μέρος του περιθωρίου.**
mai	Παίρνει ως τιμές ένα αριθμητικό διάνυσμα, στο οποίο δηλώνεται το μέγεθος του περιθωρίου προσδιορισμένο σε ίντσες.
pin	Παίρνει ως τιμές ένα αριθμητικό διάνυσμα, στο οποίο δηλώνεται το μέγεθος του διαγράμματος (πλάτος, ύψος).

Για περισσότερες πληροφορίες σχετικά με ορίσματα που αφορούν την εικόνα ενός διαγράμματος, πληκτρολογήστε help("par").

Γραμματοσειρές

```
> plot(1:10,1:10,type="n")
```

```
>windowsFonts(
A=windowsFont("Arial Black"),
B=windowsFont("Bookman Old Style"),
C=windowsFont("Times New Roman"))
> text(3,3,"Statistics - Default")
> text(4,4, family="A", "Statistics -
Arial Black")
> text(5,5, family="B", "Statistics -
Bookman Old Style")
> text(6,6, family="C", "Statistics -
Times New Roman")
```



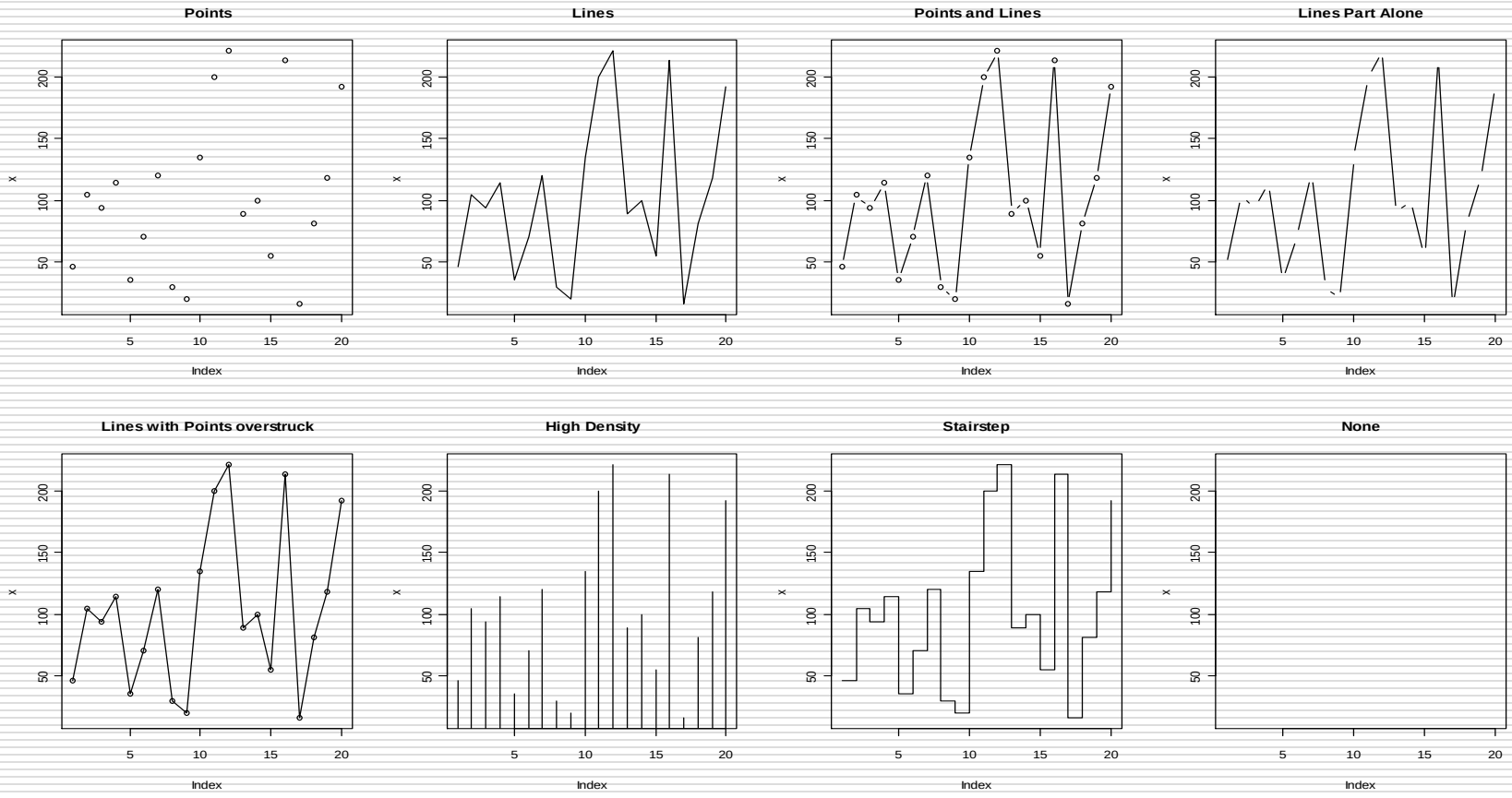
Πολλαπλά Διαγράμματα

- Μπορούμε στο ίδιο παράθυρο να έχουμε πολλά διαγράμματα σε μια διάταξη με γραμμές και στήλες, με τη βοήθεια των εντολών
  `par(mfrow=c(n,m))` ή
  `par(mfcol=c(n,m))`

Πολλαπλά Διαγράμματα

```
> par(mfrow=c(2,4))
> plot(x, type="p")
> title(main="Points")
> plot(x, type="l")
> title(main="Lines")
> plot(x, type="b")
> title(main="Points and
    Lines")
> plot(x, type="c")
> title(main="Lines Part
    Alone")
> plot(x, type="o")
> title(main="Lines with
    Points overstruck")
> plot(x, type="h")
> title(main="High
    Density")
> plot(x, type="s")
> title(main="Stairstep")
> plot(x, type="n")
> title(main="None")
```

Πολλαπλά Διαγράμματα



Πολλαπλά Διαγράμματα

- Σε ένα γράφημα μπορούμε να προσθέσουμε διάφορες γραμμές οι οποίες μπορούν να είναι διαφορετικού είδους (lty) ή χρώματος (col) για να τις διαφοροποιήσουμε με τις εντολές abline και line.

Πολλαπλά Διαγράμματα

```
> plot(x, ylim=c(0,250))
```

```
> abline(v=10, col=2)
```
→ Η ευθεία x=10

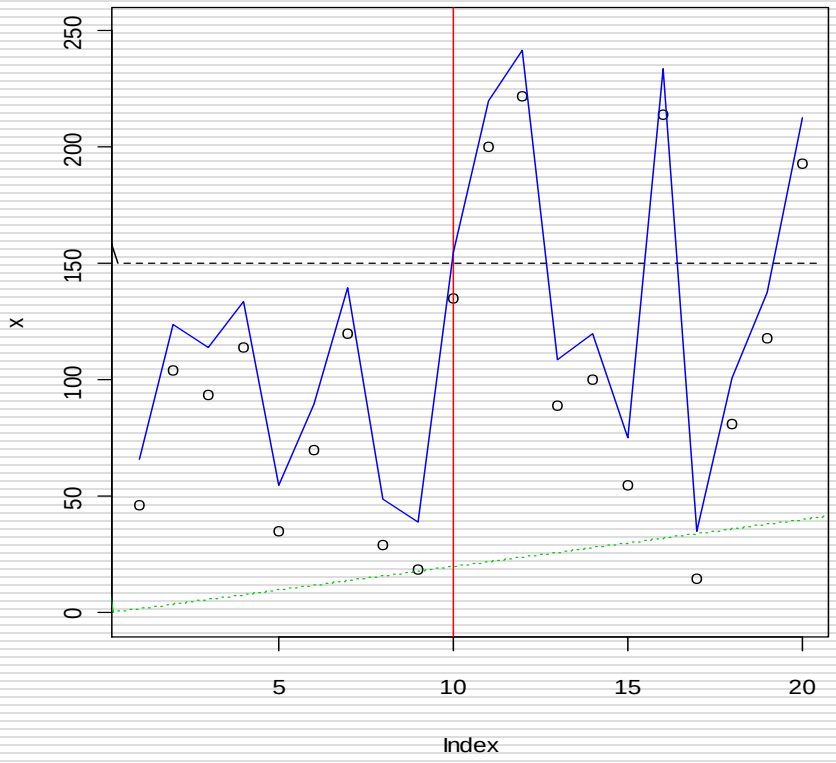
```
> abline(h=150, lty=2)
```
→ Η ευθεία y=150

```
> abline(0,2, lty=3, col=3)
```
→ Η ευθεία y=0 + 2x

```
> y<-x+20
```
→ Νέα δεδομένα y

```
> lines(y, col=4)
```
→ Απεικόνιση των y τα οποία είναι ενωμένα με ευθεία.

Πολλαπλά Διαγράμματα



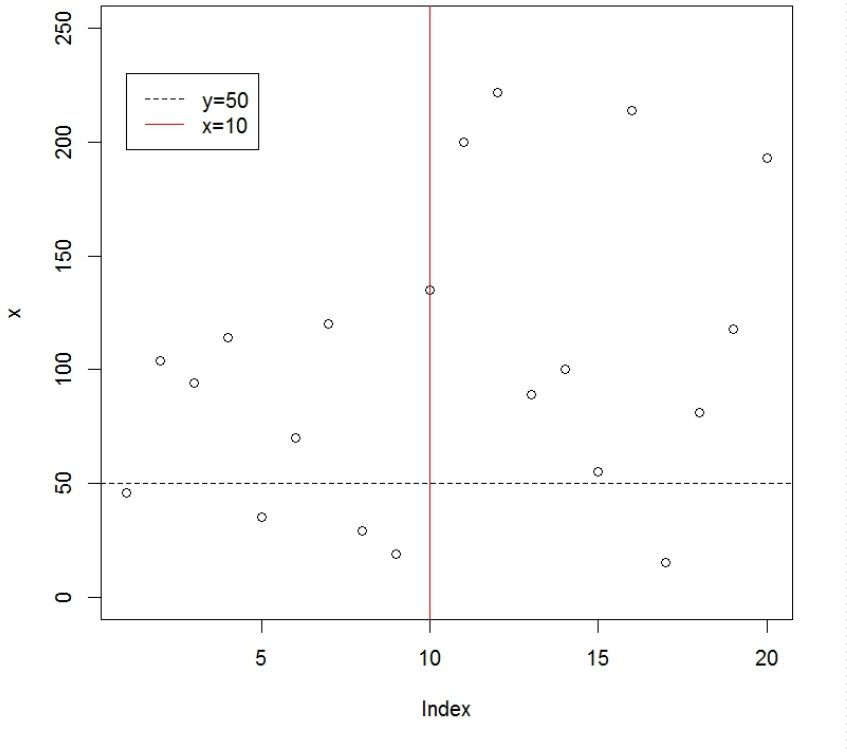
Λεζάντα

- Μπορούμε να προσθέσουμε μία λεζάντα σε ένα διάγραμμα με τη βοήθεια της εντολής legend().
- Βασικά Ορίσματα
  - (x,y): τις συντεταγμένες
  - lty: τον τύπο της γραμμής
  - col: το χρώμα για κάθε στοιχείο της λεζάντας
  - legend: το διάνυσμα χαρακτήρων ή το συνοδευτικό κείμενο

Λεζάντα

```
> x<-
c(46,104,94,114,35,70,120
,29,19,135,200,222,89,100
,55,214,15,81,118,193)

> plot(x, ylim=c(0,250))
> abline(v=10, col="red")
> abline(h=50,lty=2)
> legend(1, 230,
lty=c(2,1), col=1:2,
legend=c("y=50","x=10") )
```



Λεζάντα

- Η θέση της λεζάντας μπορεί επίσης να καθοριστεί αντικαθιστώντας τις συντεταγμένες με τις ακόλουθες λέξεις (σε εισαγωγικά):
- “bottomright”, “bottom”, “bottomleft”, “left”, “topleft”, “top”, “topright”, “right”, “center”.

Πολλαπλά Διαγράμματα

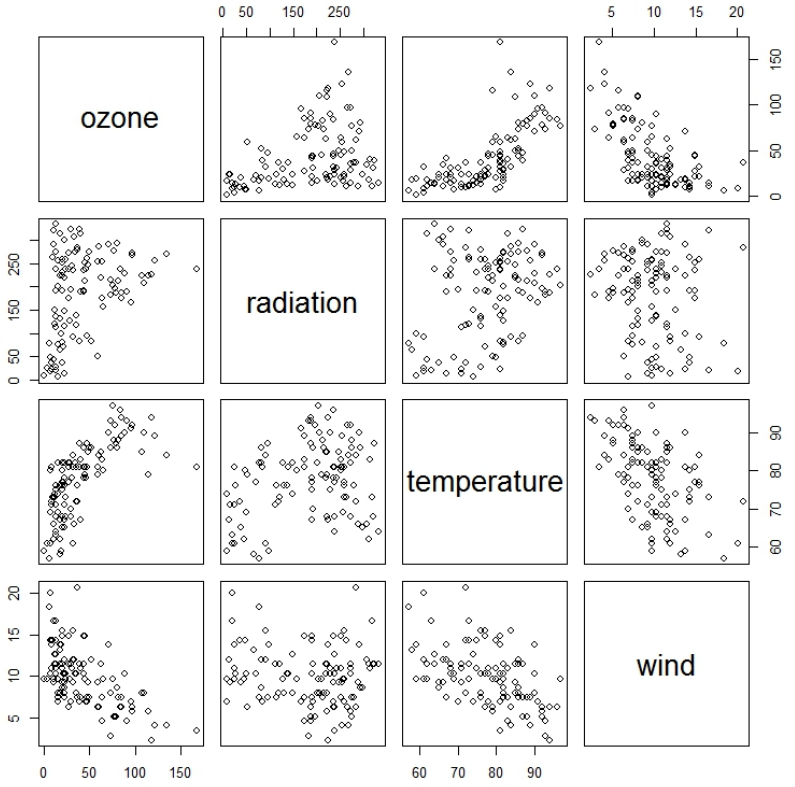
- Πολλαπλά διαγράμματα μπορούμε να δημιουργήσουμε και με χρήση της εντολής layout().

Διαγράμματα σε Μεγαλύτερες Διαστάσεις

- Στην περίπτωση που διαθέτουμε πληροφορία για περισσότερες των δύο μεταβλητών μπορούμε και πάλι να διερευνήσουμε πιθανή συσχέτιση χρησιμοποιώντας την εντολή plot() για κάθε ζεύγος μεταβλητών χωριστά. Ωστόσο, αυτό γίνεται αυτόματα με την εντολή pairs().

Διαγράμματα σε Μεγαλύτερες Διαστάσεις

```
> library(ElemStatLearn)
> data(ozone)
> pairs(ozone)
```

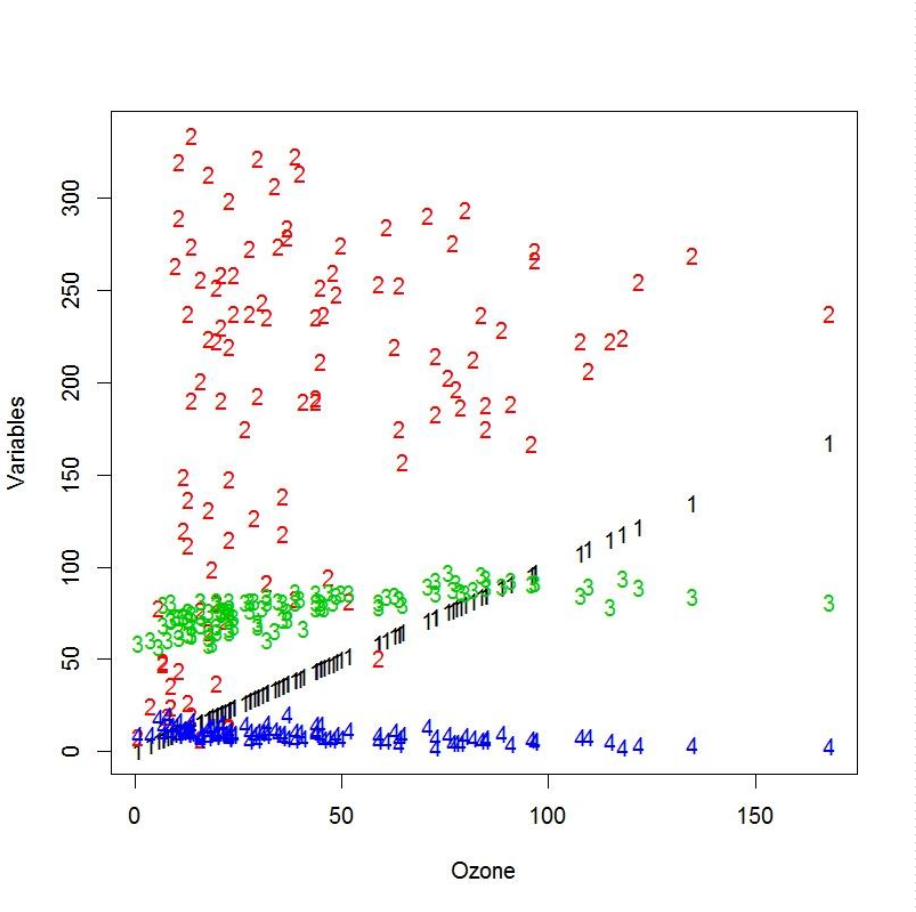


Διαγράμματα σε Μεγαλύτερες Διαστάσεις

- Εναλλακτικά, αν μας ενδιαφέρει να σχεδιάσουμε το διάγραμμα διασποράς μίας συγκεκριμένης τυχαίας μεταβλητής συναρτήσει των υπολοίπων, μπορούμε να χρησιμοποίησουμε την εντολή matplot().

Διαγράμματα σε Μεγαλύτερες Διαστάσεις

```
> matplot(ozone$ozone,
    ozone, xlab="Ozone",
    ylab="Variables")
```

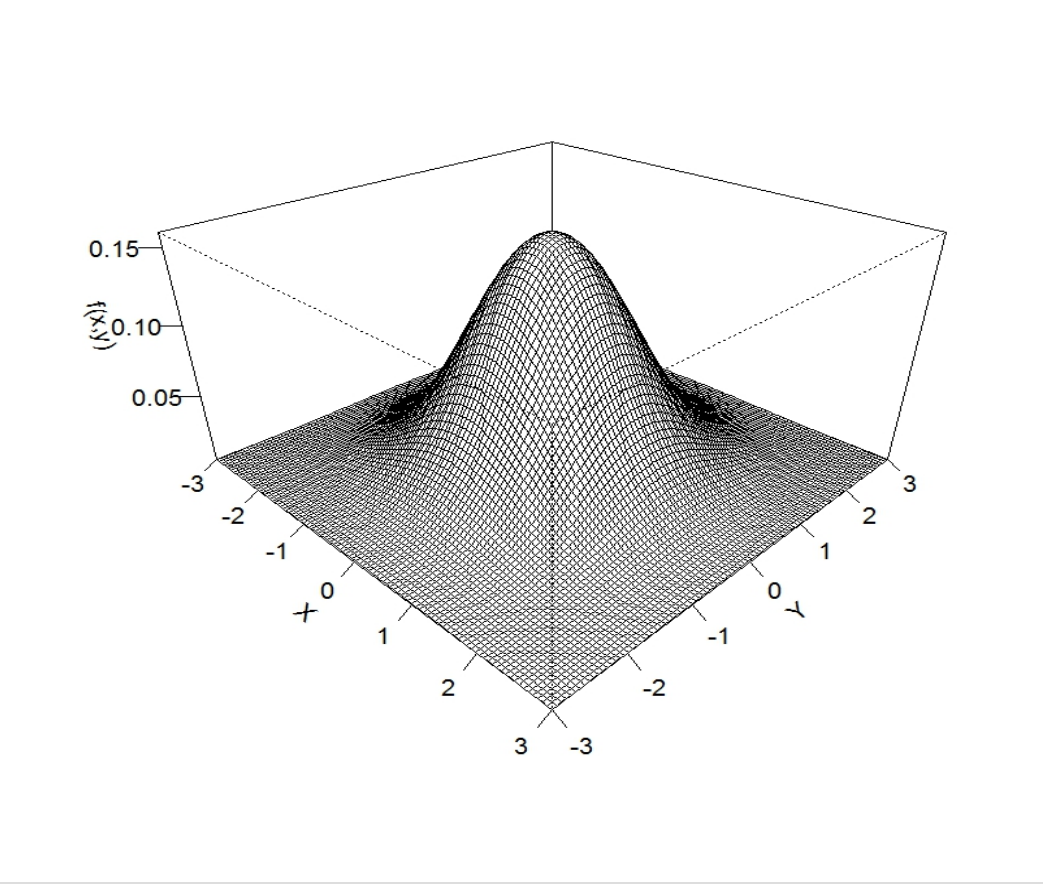


Διαγράμματα σε Μεγαλύτερες Διαστάσεις

- Προοπτική απεικόνιση με χρήση της εντολής persp().

```
> f<-function(x,y){  # the standard bivariate normal density
  z<-(1/(2*pi))*exp(-0.5*(x^2+y^2))
  }
  > y<-seq(-3,3, length=100)
  > x<-seq(-3,3, length=100)
  > z<-outer(x,y,f)  # compute density for all (x,y)
  > persp(x,y,z, theta=45, phi=30, expand=0.6,
  ticktype="detailed", xlab="X", ylab="Y", zlab="f(x,y)")
```

Διαγράμματα σε Μεγαλύτερες Διαστάσεις

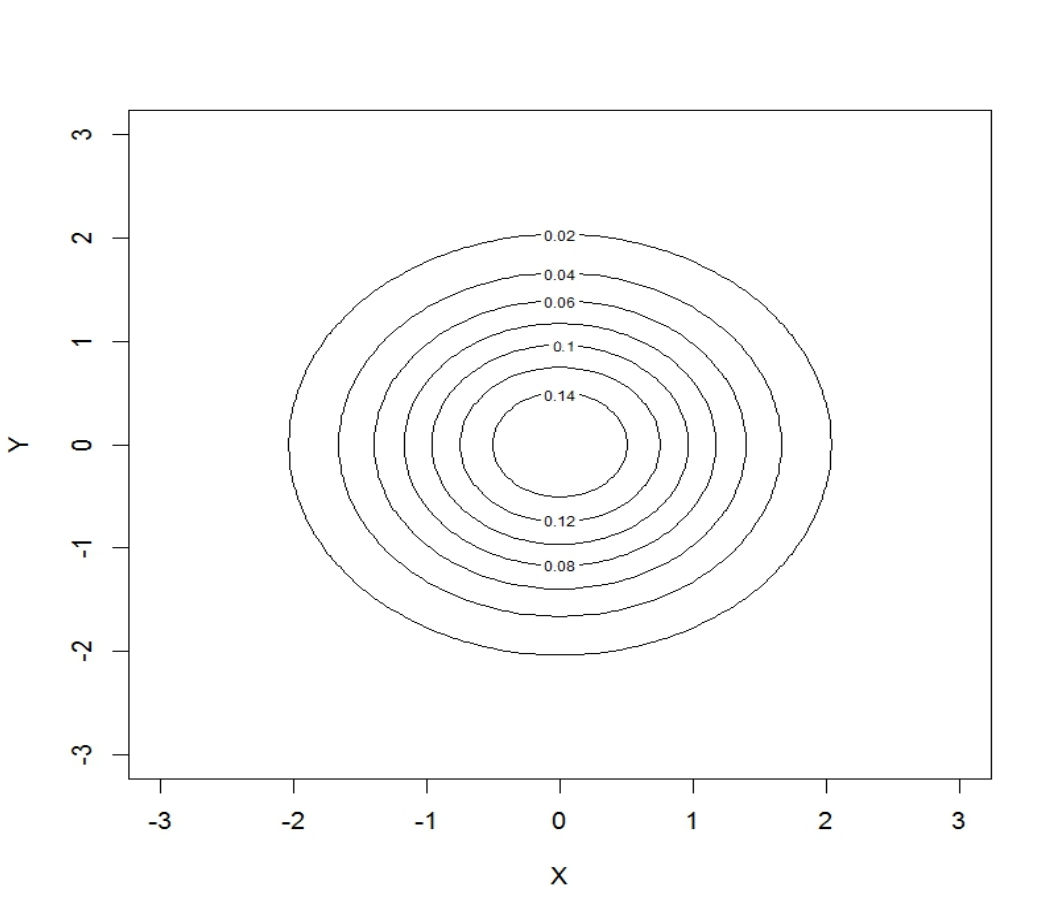


Διαγράμματα σε Μεγαλύτερες Διαστάσεις

- Διάγραμμα ισοϋψών καμπυλών με χρήση της εντολής contour().
  - Για το διάγραμμα της δισδιάστατης Τυποποιημένης Κανονικής κατανομής εκτελούμε την παρακάτω εντολή:

```
> contour(x,y,z, xlab="X", ylab="Y")
```

Διαγράμματα σε Μεγαλύτερες Διαστάσεις

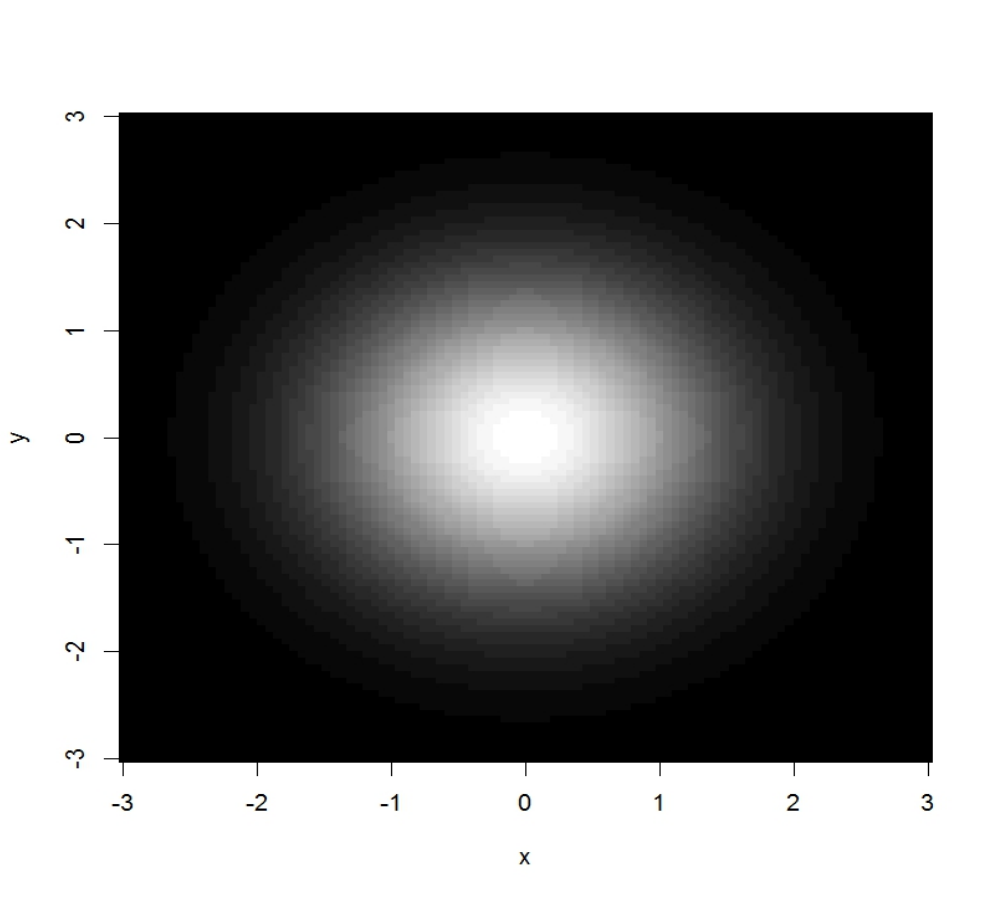


Διαγράμματα σε Μεγαλύτερες Διαστάσεις

- Έγχρωμο γραφικό πλέγμα με χρήση της εντολής image().
  - Για το διάγραμμα της δισδιάστατης Τυποποιημένης Κανονικής κατανομής εκτελούμε την παρακάτω εντολή:

```
> image(x,y,z, col=gray((0:32)/32))
```

Διαγράμματα σε Μεγαλύτερες Διαστάσεις



Αποθήκευση Διαγράμματος σε Αρχείο

- Σε Windows
  - Με χρήση του μενού File της R, όπου διαλέγετε από τις διαθέσιμες επιλογές την μορφή αποθήκευσης.
  - Κάνοντας δεξί κλικ πάνω στο διάγραμμα διαλέγετε Save as metafile…. Ή Save as postscript….
- Σε Mac
  - Με χρήση του μενού File, όπου το διάγραμμα αποθηκεύεται σε μορφή pdf.

    (Ανοίγοντας το αρχείο εν συνεχεία από το μενού File επιλέγετε την εντολή Save as και μετατρέπετε το αρχείο στη μορφή που επιθυμείτε.)

Αποθήκευση Διαγράμματος σε Αρχείο

- Μπορούμε να χρησιμοποιήσουμε κατάλληλες συναρτήσεις που δηλώνουν την μορφή του αρχείου που θέλουμε να αποθηκευτεί ένα διάγραμμα σε οποιοδήποτε λειτουργικό (υπολογιστή).
- Π.χ. αν θέλουμε να δημιουργήσουμε ένα JPG αρχείο, ονόματι plot.jpg χρησιμοποιούμε τις εξής εντολές

```
> jpeg("plot.jpeg")
> plot(x,y)
> title(main="Scatterplot", sub="30/08/2013", xlab="X", ylab="Y")
> dev.off()
```

Αποθήκευση Διαγράμματος σε Αρχείο

Μορφή	Συνάρτηση	Παρατηρήσεις
JPG	jpeg()	Μπορεί να χρησιμοποιηθεί παντού, αλλά δεν μπορείτε να αλλάξετε μέγεθος.
PNG	png()	Μπορεί να χρησιμοποιηθεί παντού, αλλά δεν μπορείτε να αλλάξετε μέγεθος.
WMF	win.metafile()	Για windows λειτουργικό. Η καλύτερη επιλογή για word. Εύκολα αλλάζετε μέγεθος.
PDF	pdf()	Η καλύτερη επιλογή για pdflatex. Εύκολα αλλάζετε μέγεθος.
Postscript	postscript()	Η καλύτερη επιλογή για latex και open office. Εύκολα αλλάζετε μέγεθος.