Ταξινόμηση Ινδών σε μια από τρεις φυλές, Brahim (Π1), Artisan (Π2), Korwa (Π3).

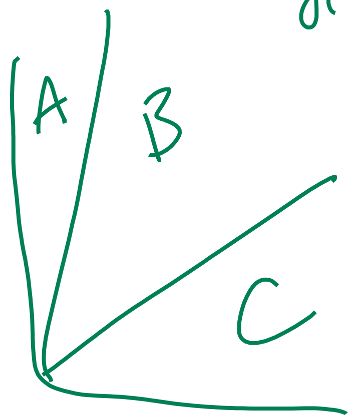
Μεταβλητές: X1=stature, X2=sitting height, X3=nasal depth, X4=nasal height.

Διαχωρισμό του χώρου που ορίζουν οι 4 μεταβλητές σε 3 περιοχές ταξινόμησης.
Ισα κόστη εσφαλμένης ταξινόμησης.

Ισες εκ των προτέρων πιθανότητες ταξινόμησης $p1 = p2 = p3 = 1/3$ 3 κατηγορίες

Δίδονται οι μέσοι, οι τυπικές αποκλίσεις και ο Πίνακας Συσχετίσεων.

2 κατηγορίες



```
> ################# # dianysmata meswn timwn
> mu1<-c(164.51,86.43,25.49,51.24)
> mu2<-c(160.53,81.47,23.84,48.62)
> mu3<-c(158.17,81.16,21.44,46.72)
> ################# # typikes apokliseis
> s1<-5.74
> s2<-3.20
> s3<-1.75
> s4<-3.50
> ################### # diakymanseis
> s11<-s1^2
> s22<-s2^2
> s33<-s3^2
> s44<-s4^2
> ################### Ypologismos stoixeiwn Pinaka S
```

των (X_1, X_2, X_3, X_4)

των X_1, X_2, X_3, X_4

→ όλα τα X_1 όλων των πληθυσμών

→ " " X_2 "

→ " " X_3 "

→ " " X_4 "

— διασπορά X_1, X_2, X_3, X_4

οι 3 πληθυσμοί έχουν ΚΟΙΝΗ ΔΙΑΣΠΟΡΑ

```
> s12<-0.5849*s1*s2
> s12
[1] 10.74344
> s13<-0.1774*s1*s3
> s13
[1] 1.781983
> s14<-0.1974*s1*s4
> s14
[1] 3.965766
> s23<-0.2094*s2*s3
> s23
[1] 1.17264
> s24<-0.2170*s2*s4
> s24
[1] 2.4304
> s34<-0.2910*s3*s4
> s34
[1] 1.782375
```

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var X \cdot Var Y}} \Rightarrow$$

$$\boxed{Cov(X,Y) = \rho(X,Y)\sqrt{Var X}\cdot\sqrt{Var Y}}$$

ex $S_{12} = Cov(X_1, X_2) = \rho_{12} * S_1 * S_2$

$$S = \begin{pmatrix} Var X_1 & S_{12} & S_{13} & S_{14} \\ S_{21} & Var X_2 & S_{23} & S_{24} \\ S_{31} & & Var X_3 & S_{34} \\ S_{41} & & & Var X_4 \end{pmatrix}$$

S_1 $=S_{22}$ S_{33} S_{44}

```
> ################ Dimiourgia Pinaka S
> c1<-c(s11,s12,s13,s14)
> c2<-c(s12,s22,s23,s24)
> c3<-c(s13,s23,s33,s34)
> c4<-c(s14,s24,s34,s44)
> S<-rbind(c1,c2,c3,c4)
> corre<-cov2cor(S)
> corre
     [,1]   [,2]   [,3]   [,4]
c1 1.0000 0.5849 0.1774 0.1974
c2 0.5849 1.0000 0.2094 0.2170
c3 0.1774 0.2094 1.0000 0.2910
c4 0.1974 0.2170 0.2910 1.0000
```

```
> S # o pinakas
      [,1]      [,2]      [,3]      [,4]
c1 32.947600 10.74344 1.781983  3.965766
c2 10.743443 10.24000 1.172640  2.430400
c3  1.781983  1.17264 3.062500  1.782375
c4  3.965766  2.43040 1.782375 12.250000
> eigen(S)$values # ypologismos idiotimwn tou S ###############
[1] 38.200422 11.774581  5.868467  2.656629
```

S = Πίνακας διασπορών – συνδιασπορών

ΣΥΜΜΕΤΡΙΚΟΣ ⇒ (διαγωνοποιείται)

```
> # antistrofos tou S
> Sinv<-solve(S)
> Sinv
              c1          c2           c3           c4
[1,]  0.046610571 -0.04703200 -0.006294337 -0.004842544
[2,] -0.047032003  0.15251569 -0.024344286 -0.011491066
[3,] -0.006294337 -0.02434429  0.366558171 -0.046466604
[4,] -0.004842544 -0.01149107 -0.046466604  0.092241077
```

αυτόνομος όρος (διάλεξη 11)
είναι το 'b'!

```
> ### Ypologismos gia 1,2
> round(Sinv%*%(mu1-mu2),digits=4)
      [,1]
[1,] -0.0708
[2,]  0.4990
[3,]  0.3373
[4,]  0.0887
> 0.5*(mu1+mu2)%*%Sinv%*%(mu1-mu2)
       [,1]
[1,] 43.12859
> ### Ypologismos  gia 13
> round(Sinv%*%(mu1-mu3),digits=4)
     [,1]
[1,] 0.0003
[2,] 0.3550
```

```
[3,] 1.1063
[4,] 0.1375
> 0.5*(mu1+mu3)%*%Sinv%*%(mu1-mu3)
      [,1]
[1,] 62.48837
> ### Ypologismos gia  23
> round(Sinv%*%(mu2-mu3),digits=4)
      [,1]
[1,]  0.0711
[2,] -0.1440
[3,]  0.7691
[4,]  0.0487
> 0.5*(mu2+mu3)%*%Sinv%*%(mu2-mu3)
      [,1]
[1,] 19.35978
```

ℓ

$$x \cdot \ell > \frac{1}{2}\bar{x}_1 \ell + \frac{1}{2}\bar{x}_2 \ell$$

$$x \ell - \frac{1}{2}\boxed{\bar{x}_1}\ell - \frac{1}{2}\boxed{\bar{x}_2}\ell > 0$$

$$\left[x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right]\ell > 0$$

```
> (mu2-0.5*(mu1+mu2))%*%Sinv%*%(mu2-mu1)
      [,1]
[1,] 1.491079
> (mu3-0.5*(mu1+mu3))%*%Sinv%*%(mu3-mu1)
      [,1]
[1,] 3.487421
> (mu3-0.5*(mu2+mu3))%*%Sinv%*%(mu3-mu2)
      [,1]
[1,] 1.03077
> (mu1-0.5*(mu1+mu2))%*%Sinv%*%(mu1-mu2)
      [,1]
[1,] 1.491079
> (mu1-0.5*(mu1+mu3))%*%Sinv%*%(mu1-mu3)
      [,1]
[1,] 3.487421
```

Αν εφαρμοσθεί κάποιος με $x_1, x_2, x_3, x_4 \equiv$ mu2 που θα χρησιμοποιηθεί;

NEW OBSERVATION m4

απέχει αρκετά από το Π_1. Που τοποθετείται?

```
> mu4<-c(161,81.16,23.44,49.72)
> (mu4-0.5*(mu1+mu2))%*%Sinv%*%(mu2-mu1)
         [,1]
[1,] 1.716376
> (mu4-0.5*(mu1+mu3))%*%Sinv%*%(mu3-mu1)
          [,1]
[1,] 0.8615478
> (mu4-0.5*(mu2+mu3))%*%Sinv%*%(mu3-mu2)
           [,1]
[1,] -0.8548277
```

$\Rightarrow$ mu4 $\in \Pi_2$

mu4 $\in \Pi_3$

$\Leftarrow$ συγκρίνω Π_2 & Π_3

Άρα ανήκει Π_2

```
> #####use the lda() function from the MASS package to fit the LDA model
> library(MASS)
> attach(iris)
> iris
```

	Sepal.L	Sepal.W	Petal.L	Petal.W	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa

R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. 7 (2): 179–188.

15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
32	5.4	3.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa

36	5.0	3.2	1.2	0.2	setosa
37	5.5	3.5	1.3	0.2	setosa
38	4.9	3.6	1.4	0.1	setosa
39	4.4	3.0	1.3	0.2	setosa
40	5.1	3.4	1.5	0.2	setosa
41	5.0	3.5	1.3	0.3	setosa
42	4.5	2.3	1.3	0.3	setosa
43	4.4	3.2	1.3	0.2	setosa
44	5.0	3.5	1.6	0.6	setosa
45	5.1	3.8	1.9	0.4	setosa
46	4.8	3.0	1.4	0.3	setosa
47	5.1	3.8	1.6	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
49	5.3	3.7	1.5	0.2	setosa
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	**versicolor**
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor

57	6.3	3.3	4.7	1.6 versicolor
58	4.9	2.4	3.3	1.0 versicolor
59	6.6	2.9	4.6	1.3 versicolor
60	5.2	2.7	3.9	1.4 versicolor
61	5.0	2.0	3.5	1.0 versicolor
62	5.9	3.0	4.2	1.5 versicolor
63	6.0	2.2	4.0	1.0 versicolor
64	6.1	2.9	4.7	1.4 versicolor
65	5.6	2.9	3.6	1.3 versicolor
66	6.7	3.1	4.4	1.4 versicolor
67	5.6	3.0	4.5	1.5 versicolor
68	5.8	2.7	4.1	1.0 versicolor
69	6.2	2.2	4.5	1.5 versicolor
70	5.6	2.5	3.9	1.1 versicolor
71	5.9	3.2	4.8	1.8 versicolor
72	6.1	2.8	4.0	1.3 versicolor
73	6.3	2.5	4.9	1.5 versicolor
74	6.1	2.8	4.7	1.2 versicolor
75	6.4	2.9	4.3	1.3 versicolor
76	6.6	3.0	4.4	1.4 versicolor
77	6.8	2.8	4.8	1.4 versicolor

78	6.7	3.0	5.0	1.7 versicolor
79	6.0	2.9	4.5	1.5 versicolor
80	5.7	2.6	3.5	1.0 versicolor
81	5.5	2.4	3.8	1.1 versicolor
82	5.5	2.4	3.7	1.0 versicolor
83	5.8	2.7	3.9	1.2 versicolor
84	6.0	2.7	5.1	1.6 versicolor
85	5.4	3.0	4.5	1.5 versicolor
86	6.0	3.4	4.5	1.6 versicolor
87	6.7	3.1	4.7	1.5 versicolor
88	6.3	2.3	4.4	1.3 versicolor
89	5.6	3.0	4.1	1.3 versicolor
90	5.5	2.5	4.0	1.3 versicolor
91	5.5	2.6	4.4	1.2 versicolor
92	6.1	3.0	4.6	1.4 versicolor
93	5.8	2.6	4.0	1.2 versicolor
94	5.0	2.3	3.3	1.0 versicolor
95	5.6	2.7	4.2	1.3 versicolor
96	5.7	3.0	4.2	1.2 versicolor
97	5.7	2.9	4.2	1.3 versicolor
98	6.2	2.9	4.3	1.3 versicolor

99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	**virginica**
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica
106	7.6	3.0	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica
110	7.2	3.6	6.1	2.5	virginica
111	6.5	3.2	5.1	2.0	virginica
112	6.4	2.7	5.3	1.9	virginica
113	6.8	3.0	5.5	2.1	virginica
114	5.7	2.5	5.0	2.0	virginica
115	5.8	2.8	5.1	2.4	virginica
116	6.4	3.2	5.3	2.3	virginica
117	6.5	3.0	5.5	1.8	virginica
118	7.7	3.8	6.7	2.2	virginica
119	7.7	2.6	6.9	2.3	virginica

120	6.0	2.2	5.0	1.5	virginica
121	6.9	3.2	5.7	2.3	virginica
122	5.6	2.8	4.9	2.0	virginica
123	7.7	2.8	6.7	2.0	virginica
124	6.3	2.7	4.9	1.8	virginica
125	6.7	3.3	5.7	2.1	virginica
126	7.2	3.2	6.0	1.8	virginica
127	6.2	2.8	4.8	1.8	virginica
128	6.1	3.0	4.9	1.8	virginica
129	6.4	2.8	5.6	2.1	virginica
130	7.2	3.0	5.8	1.6	virginica
131	7.4	2.8	6.1	1.9	virginica
132	7.9	3.8	6.4	2.0	virginica
133	6.4	2.8	5.6	2.2	virginica
134	6.3	2.8	5.1	1.5	virginica
135	6.1	2.6	5.6	1.4	virginica
136	7.7	3.0	6.1	2.3	virginica
137	6.3	3.4	5.6	2.4	virginica
138	6.4	3.1	5.5	1.8	virginica
139	6.0	3.0	4.8	1.8	virginica
140	6.9	3.1	5.4	2.1	virginica

141	6.7	3.1	5.6	2.4	virginica
142	6.9	3.1	5.1	2.3	virginica
143	5.8	2.7	5.1	1.9	virginica
144	6.8	3.2	5.9	2.3	virginica
145	6.7	3.3	5.7	2.5	virginica
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

```
> fit.iris<-lda(Species~., data=iris)
> fit.iris
Call:
lda(Species ~ ., data = iris)
```

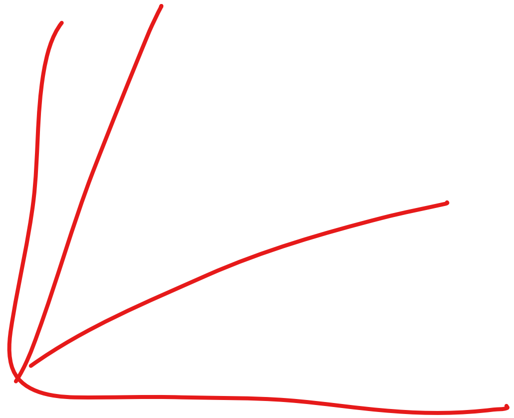
Prior probabilities of groups:

```
   setosa versicolor  virginica
0.3333333  0.3333333  0.3333333
```

$p_1 = p_2 = p_3 = \frac{50}{150} = \frac{1}{3}$

Group means:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026



Coefficients of linear discriminants:

	LD1	LD2
Sepal.Length	0.8293776	0.02410215
Sepal.Width	1.5344731	2.16452123
Petal.Length	-2.2012117	-0.93192121
Petal.Width	-2.8104603	2.83918785

Proportion of trace:

LD1	LD2
0.9912	0.0088

> ####Proportion of trace: The % of separation achieved by each linear discriminant function.

- **LD1:** .8293*Sepal.Length + 1.53*Sepal.Width – 2.2012*Petal.Length – 2.8104*Petal.Width
- **LD2:** .024*Sepal.Length + 2.164*Sepal.Width – 0.931*Petal.Length + 2.839*Petal.Width

#PREDICTIONS

```
predicted <- predict(fit.iris, iris)
predicted$class
```

#find accuracy of model

```
mean(predicted$class==iris$Species)
```

% accuracy $= \frac{147}{150} = \frac{150-3}{150}$

$= 0.98$

#PLOT

```
library(ggplot2)
plot1 <- cbind(iris, predict(fit.iris)$x)
```

#create plot

```
ggplot(plot1, aes(LD1, LD2)) +
  geom_point(aes(color = Species))
```

τυποποίηση ⇒ μέσος = 0, διασπορά = 1

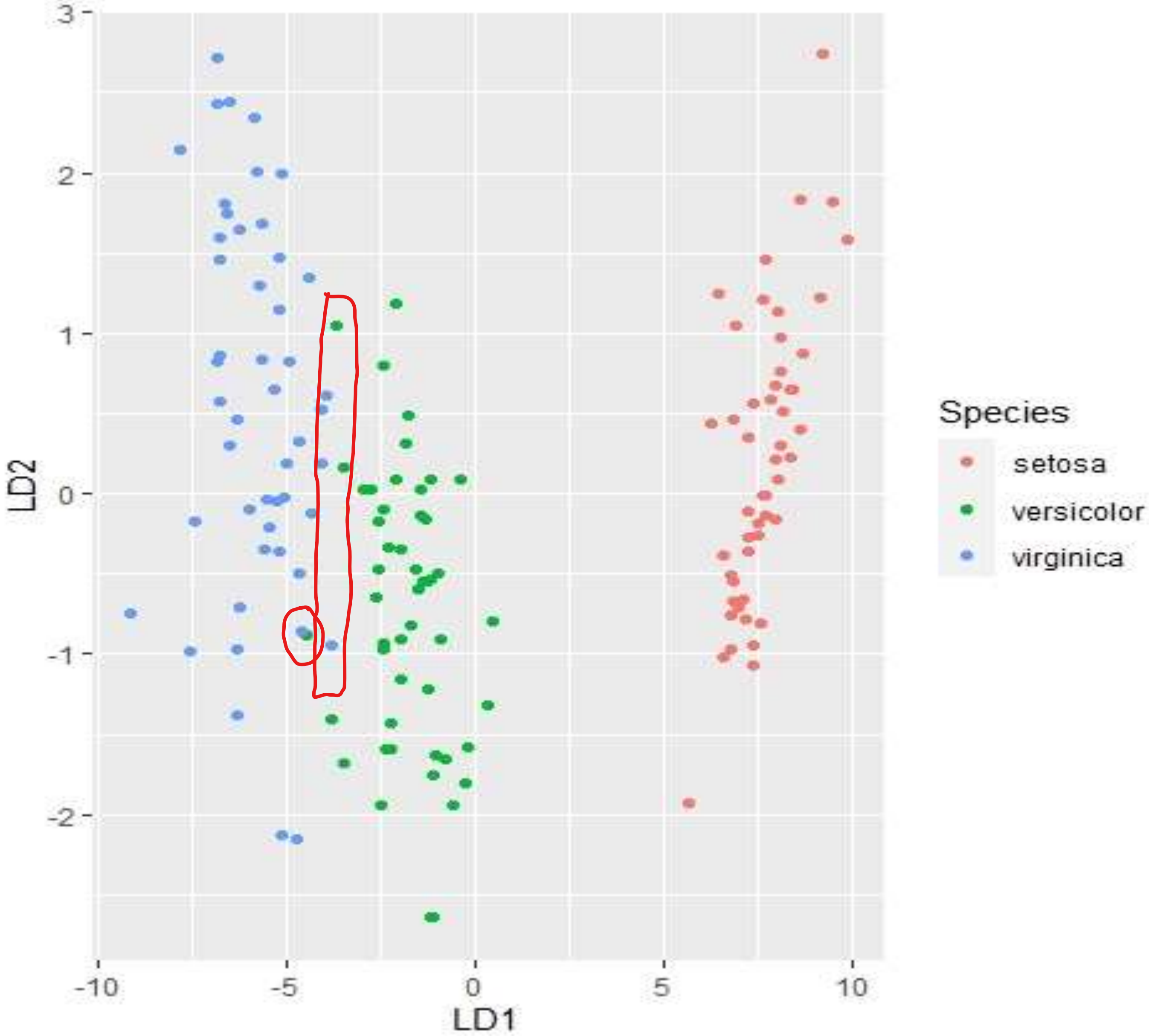
#Each predictor has the same variance> NO, use "scale"

```
 #scale each predictor variable (i.e. first 4 columns)
# iris[1:4] <- scale(iris[1:4])
```

Να χρησιμοποιηθει το 65% των data ως TRAINING SET και να ελεγχθει η αποτελεσαμτικοτητα του στο υπολοιπο 35% TESTING SET.

δηλαδή 100 από τα 150

τα υπόλοιπα 50 λέγονται αυτά έχει του πρέπει



3
2
1
0
-1
-2
LD2
-10
-5
0
5
10
LD1
Species
setosa
versicolor
virginica