ΔΠΜΣ Βιοστατιστικής

και Επιστήμης Δεδομένων Υγείας

Πέτρος Τζαβέλλας

ΑΜ: 7450022400026

2η Εργασία στη Πολυμεταβλητή Στατιστική

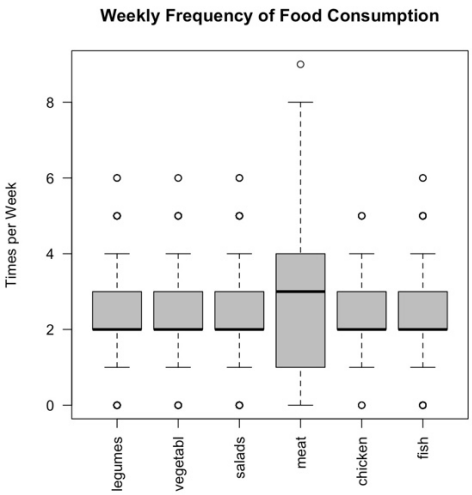
Μέρος Α:

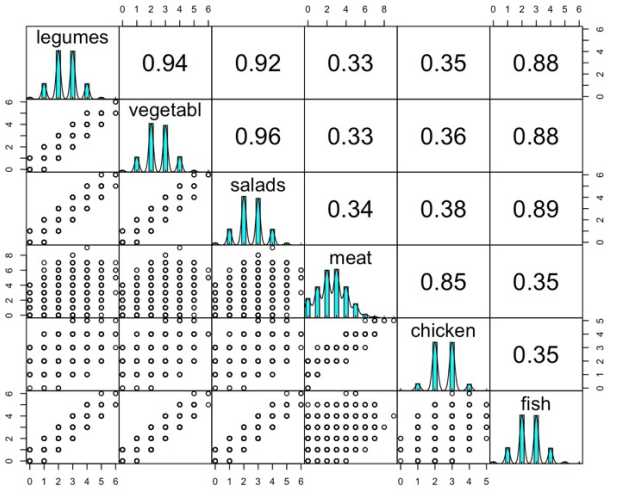
1ο Ερώτημα:

Μονομεταβλητά Περιγραφικά Στατιστικά Στοιχεία:

A/A	Legumes	Vegetables	Salads	Meat	Chicken	Fish
Median(IQR)	2(1)	2(1)	2(1)	3(3)	2(1)	2(1)
Mean(SD)	2.496(0.95)	2.496(0.94)	2.49(0.96)	2.529(1.5)	2.494(0.71)	2.493(0.94)
Variance	0.89	0.89	0.93	2.25	0.51	0.89

Οι περισσότερες μεταβλητές, πέρα από το κρέας, έχουν παρόμοια μέση τιμή (≈2.5 φορές/εβδομάδα) και χαμηλή διακύμανση. Το κρέας είναι το μόνο που έχει ασύμμετρη κατανομή και θα επιλέξουμε να το περιγράψουμε με το median(IQR). Τα όσπρια, τα λαχανικά, οι σαλάτες και τα ψάρια έχουν συγκρίσιμες συχνότητες κατανάλωσης και παρόμοιες διασπορές. Τέλος, το κοτόπουλο έχει την μικρότερη διακύμανση, ένδειξη πιο ομοιογενούς κατανάλωσης.





Πίνακας Συσχετίσεων:

	Legumes	Vegetables	Salads	Meat	Chicken	Fish
Legumes	1	0.94	0.92	0.33	0.35	0.88
Vegetables	0.94	1	0.96	0.33	0.36	0.88
Salads	0.92	0.96	1	0.34	0.38	0.89
Meat	0.33	0.33	0.34	1	0.85	0.35
Chicken	0.35	0.36	0.38	0.85	1	0.35
Fish	0.88	0.88	0.89	0.35	0.35	1

Τα λαχανικά, τα όσπρια, σαλάτες και ψάρια, έχουν μεταξύ τους μεγάλες συσχετίσεις, κάτι που υποδεικνύει ότι αυτοί που καταναλώνουν ένα από αυτά (πιο υγιεινά τρόφιμα), τείνουν να καταναλώνουν και τα υπόλοιπα. Αντίστοιχα, το κοτόπουλο με το κρέας έχουν πολύ ισχυρή θετική συσχέτιση, που δείχνει μια προτίμηση στο κρέας. Αντιθέτως, η pescatarian τροφές είναι ασθενή συσχετισμένες με το κρέας και το κοτόπουλο, κάτι που υποδηλώνει διαφορετικά διατροφικά προφίλ ατόμων.

2ο Ερώτημα:

Αρχικά, και από τον παρακάτω πίνακα του Kaiser-Meyer-Olkin καταλήγουμε, αφού ο συνολικός δείκτης είναι μεγαλύτερος του 0.8, ότι υπάρχουν εξαιρετικές συσχετίσεις. Το παραπάνω αποτέλεσμα επιβεβαιώνεται από τον έλεγχο σφαιρικότητας του Bartlett, ο οποίος βγαίνει ισχυρά μη στατιστικά σημαντικός και άρα απορρίπτεται η

H_0, άρα ο πίνακας συσχέτισης διαφέρει από τον ταυτοτικό πίνακα I_6, ως αποτέλεσμα αυτού μπορούμε να συνεχίσουμε με την παραγοντική ανάλυση.

Πίνακας KMO:

KMO	Value
Overall KMO	0.81
Legumes	0.9
Vegetables	0.82
Salads	0.84
Meat	0.61
Chicken	0.63
Fish	0.92

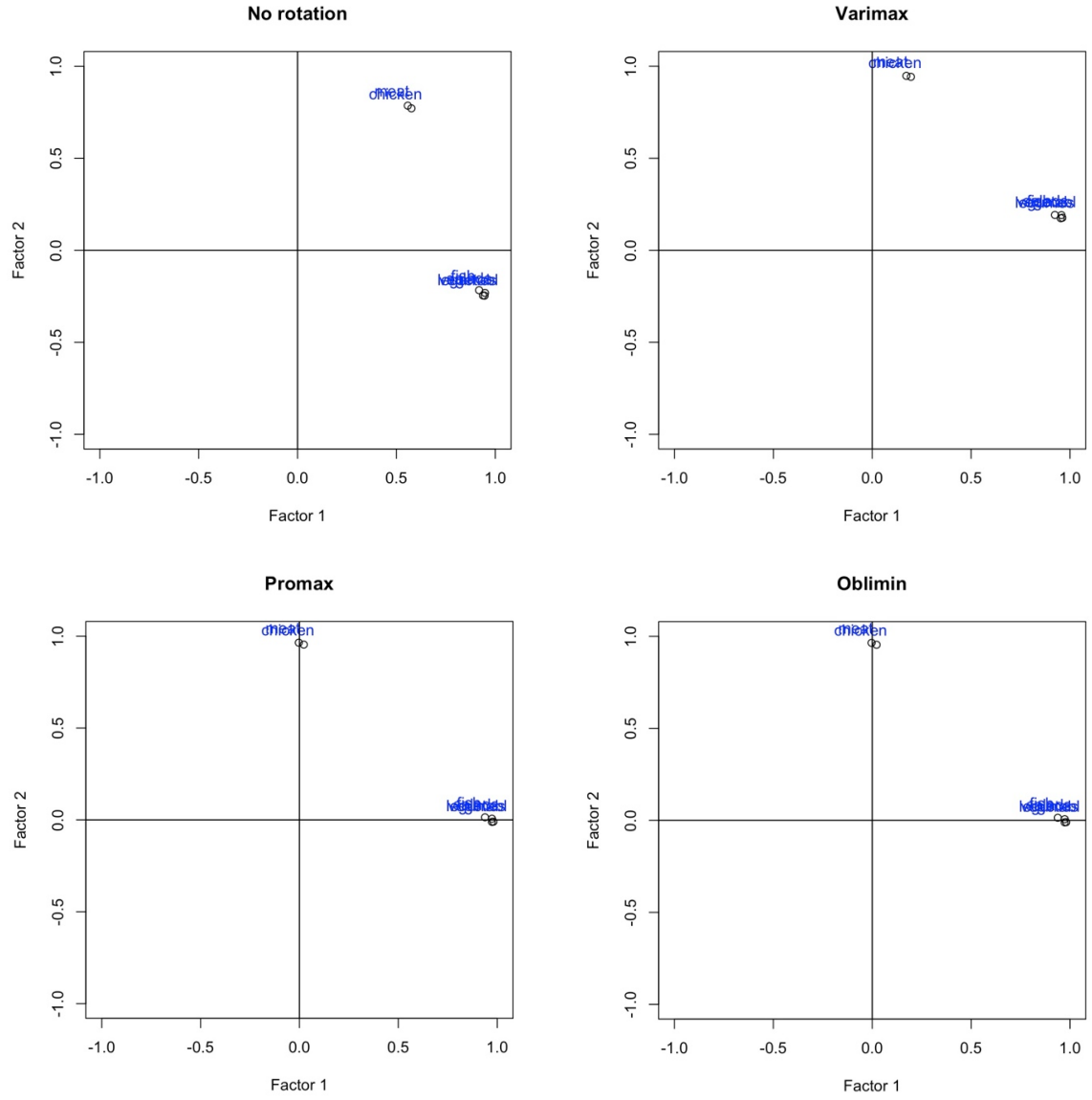
Οι κοινές διασπορές που υπολογίστηκαν από το παραγοντικό μοντέλο χωρίς στροφή είναι οι εξής που φαίνονται παρακάτω:

Πίνακας κοινών διασπορών (Communalities)

Legumes	Vegetables	Saldas	Meat	Chicken	Fish
0.939	0.956	0.952	0.928	0.926	0.89

3ο Ερώτημα:

Στον παρακάτω πίνακα διαγραμμάτων φαίνεται ανάλογα με την στροφή που επιλεγεί, το αντίστοιχο διάγραμμα φορτίων.



Αρχικά, παρατηρούμε ότι το rotation με Promax και με Oblimin δίνουν ακριβώς τα ίδια αποτελέσματα, κατ' επέκταση θα συγκρίνουμε μόνο το Promax με τις υπόλοιπες δύο στροφές.

No rotation				Varimax				Promax		
	PC1	PC2			PC1	PC2			PC1	PC2
Legumes	0.938	-0.246		Legumes	0.953	0.174		Legumes	0.973	
Vegetables	0.946	-0.247		Vegetables	0.961	0.177		Vegetables	0.981	
Salads	0.948	-0.232		Salads	0.957	0.191		Salads	0.973	
Meat	0.558	0.786		Meat	0.173	0.948		Meat		0.965
Chicken	0.576	0.771		Chicken	0.196	0.942		Chicken		0.954
Fish	0.918	-0.217		Fish	0.924	0.192		Fish	0.938	

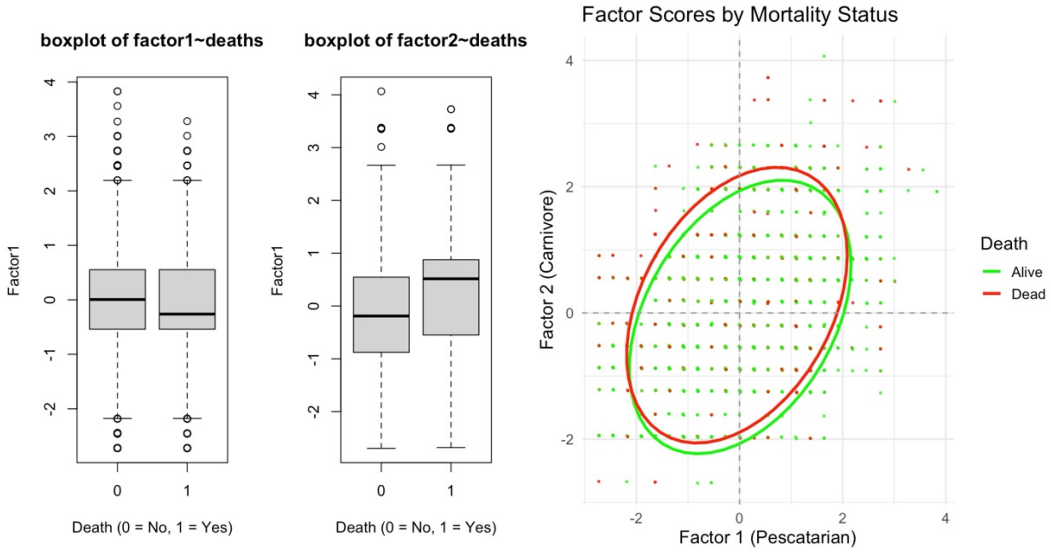
Στο παραπάνω πίνακα εμφανίζονται τα φορτία (loadings), ανάλογα το rotation του παραγοντικού μοντέλου.

Με την Varimax περιστροφή, παρατηρούμε ότι η ερμηνεία έχει διευκολυνθεί αρκετά, αφού είναι ξεκάθαρο ότι μεγάλες τιμές της πρώτης συνιστώσας συνεπάγεται και άτομο με πιο pescatarian διατροφή, ενώ μεγάλες τιμές της δεύτερης συνιστώσας συνεπάγεται σε μια πιο carnivore διατροφή.

Με την Promax περιστροφή, παρατηρούμε ότι υπάρχει η παρόμοια ερμηνεία, δηλαδή μεγάλες τιμές της πρώτης συνιστώσας συνεπάγεται μεγαλύτερη κατανάλωση λαχανικών/οσπρίων/ψαριών, με την διαφορά ότι **δεν** συνεπάγεται αναγκαστικά μικρότερη κατανάλωση κρέατος ή κοτόπουλου. Αυτό είναι αποτέλεσμα του ότι επιτρέπουμε συσχέτιση μεταξύ των δύο συνιστωσών, μιας και η περιστροφή δεν είναι ορθογώνια πλέον. Καταλήγω στην Promax περιστροφή μιας και είναι ένα πιο ρεαλιστικό σενάριο και ερμηνευτικά «μεγαλύτερη» αξία.

4ο Ερώτημα:

Παίρνοντας υπόψιν το παρακάτω γράφημα, καθώς και τις παραπάνω ερμηνείες του μοντέλου από το προηγούμενο ερώτημα, υπάρχει η ομαδοποίηση σε συμμετέχοντες που έχουν πιο κρεατοφαγικές συνήθειες (πιο πρωτεϊνική διατροφή) και συμμετέχοντες που έχουν πιο pescatarian διατροφές. Όπως φαίνεται και στο σχήμα συνδέεται οι θάνατοι με το είδος της διατροφής κάθε συμμετέχοντα, ωστόσο για να δούμε ότι αυτή η διαφορά είναι στατιστικά σημαντική θα χρειαστεί να τρέξουμε μια λογιστική παλινδρόμηση.



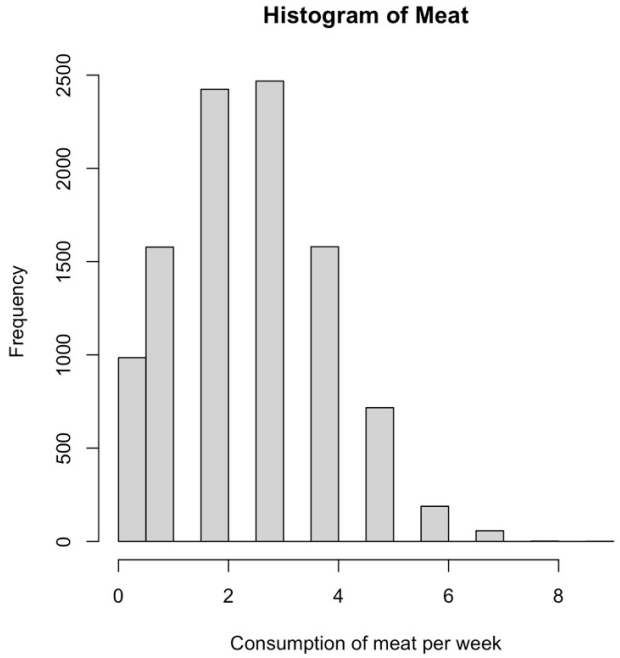
Κάνοντας παλινδρόμηση είναι ξεκάθαρο ότι ο πρώτος παράγοντας είναι προστατευτικός, δηλαδή συμμετέχοντες με pescatarian διατροφή έχουν μειωμένες

πιθανότητες κατά 15% να πεθάνουν, ενώ οι συμμετέχοντες που κάνουν πιο κρεατοφαγική διατροφή έχουν αυξημένες πιθανότητες θανάτου κατά 25% και είναι ισχυρά στατιστικά σημαντικά.

	OR	SE	z value	p-value
Component 1	0.85	0.024	-6.833	<0.0001
Component 2	1.25	0.024	9.432	<0.0001
Intercept	0.397	0.022	-41.303	<0.0001

5ο Ερώτημα:

Βλέποντας, για τις ιδιοτιμές (μόνο οι δύο πρώτες είναι μεγαλύτερες από 1), καθώς και την αθροιστική μεταβλητότητα που εξηγούν οι δύο ιδιοτιμές πρώτες ιδιοτιμές (93,2% >>80%), καταλήγουμε ότι με την μέθοδο τον κύριων συνιστωσών το μοντέλο που έχουμε καταλήξει στα παραπάνω ερωτήματα είναι το βέλτιστο με την συγκεκριμένη μέθοδο, τώρα απαιτείται να ελεγχθεί και με την μέθοδο της πιθανοφάνειας. Αρχικά το παραγοντικό μοντέλο που θα αναπτυχθεί με την μέθοδο της πιθανοφάνειας μπορεί να έχει το πολύ δύο παράγοντες. Επίσης, οι επιμέρους μεταβλητές των τροφών δεν ακολουθούν κανονική κατανομή, ενδεικτικά το παρακάτω ιστόγραμμα με το κρέας και έτσι η μέθοδος της μέγιστης πιθανοφάνειας δεν μπορεί να εφαρμοστεί. Εν κατακλείδι, καταλήγουμε στο μοντέλο των προηγούμενων ερωτημάτων.



Μέρος Β:

1ο Ερώτημα:

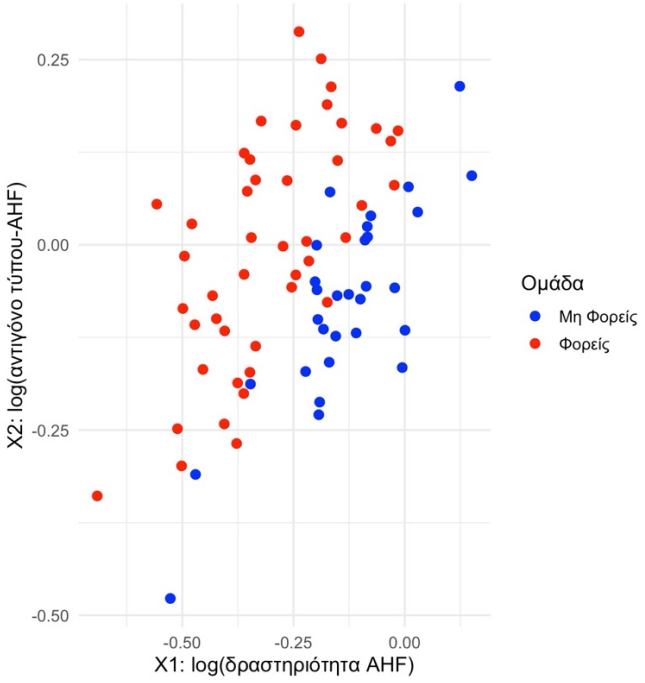
Τα κέντρα βάρους των 2 πληθυσμών είναι τα εξής m_1=(-0.135,-0.0778) και m_2=(-0.308,-0.006), επίσης οι αντίστοιχοι πίνακες συσχέτισης είναι οι εξής:

$$R_1 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \qquad R_2 = \begin{pmatrix} 1 & 0.64 \\ 0.64 & 1 \end{pmatrix}$$

Ως προς την καταλληλόλητα των δύο μεταβλητών για την διάκριση των 2 πληθυσμών θα ελέγξουμε τα εξής δύο πράγματα. Πρώτον, την διαφορά των μέσων των

μεταβλητών μεταξύ των 2 πληθυσμών, το scatterplot ώστε να παρατηρήσουμε, αν υπάρχει ξεκάθαρος διαχωρισμός. Παρατηρούμε παρακάτω ότι υπάρχει μεγάλη διαφορά μεταξύ των μέσων μεταξύ των δύο πληθυσμών, αφού $|X_{11}-X_{12}|=0.173$ και αντίστοιχα $|X_{21}-X_{22}|=0.072$ Επιπρόσθετα, στο scatterplot φαίνεται ξεκάθαρος διαχωρισμός των δύο πληθυσμών.

Μεταβλητή	Ομάδα	Μέση Τιμή	SE	Ελάχιστο	Μέγιστο
X1	1	-0.135	0.144	-0.5268	0.1507
X2	1	-0.0778	0.134	-0.4773	0.214
X1	2	-0.308	0.154	-0.6911	-0.0149
X2	2	,-0.006	0.155	-0.339	0.2876



2ο Ερώτημα:

Η διακρίνουσα συνάρτηση του Fisher που θα χρησιμοποιηθεί είναι η εξής:

$$LD=9.034666*X_1+8.003857*X_2$$

3ο Ερώτημα:

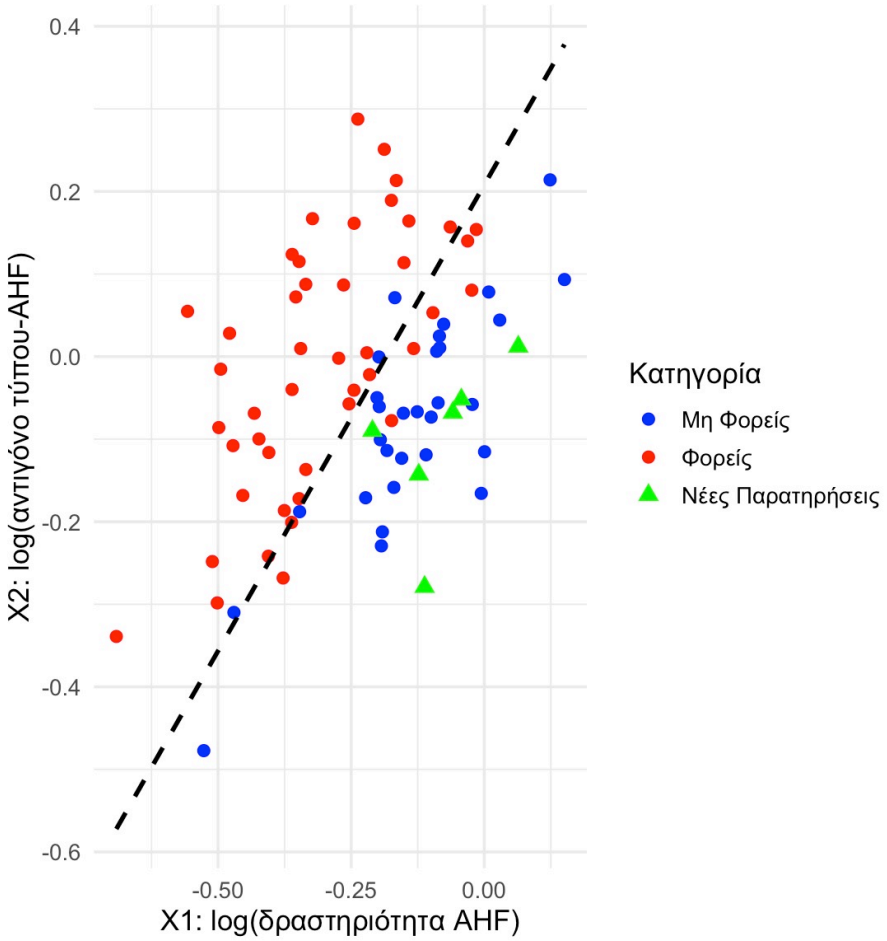
Έχοντας ταξινομήσει τις υπάρχουσες παρατηρήσεις με βάσει τον κανόνα διάκρισης του Fisher παίρνουμε τον Confusion Matrix που φαίνεται παρακάτω:

Πραγματικό/Predicted	1	2
1	27	3
2	8	37

Βάση του παραπάνω πίνακα καταλήγουμε ότι το ποσοστό δυσταξινόμησης είναι **14.67%**.

4ο Ερώτημα:

Παίρνουμε τις 6 νέες παρατηρήσεις και κάνοντας predict με βάση το κανόνα διάκρισης που έχουμε καταλήξει στα παραπάνω ερωτήματα, καταλήγουμε ότι και οι 6 αυτές παρατηρήσεις βάση του παραπάνω κανόνα θα κατέληγαν στον πρώτο πληθυσμό, των μη φορέων, όπως φαίνεται και στο παρακάτω διάγραμμα.



5ο Ερώτημα:

Αρχικά, για να ερευνήσουμε την υπόθεση διμεταβλητής κανονικότητας, θα ελέγξουμε τις περιθώριες για να δούμε, αν είναι κανονικά κατανεμημένες. Αυτή η συνθήκη είναι αναγκαία μεν, αλλά όχι ικανή. Έτσι, με το test κανονικότητας του Shapiro-Wilk, έχουν και οι δύο περιθώριες p-value>0.05 και έτσι θεωρούμε ότι και οι δύο είναι κανονικά κατανεμημένες. Στην συνέχεια, κάνουμε τον έλεγχο του Mardia για Multivariate Normal Distributions, ο οποίος βγαίνει μη στατιστικά σημαντικός και έτσι καταλήγουμε ότι η κατανομή είναι διμεταβλητή κανονική

	Test	Statistic	p.value	Method	MVN
1	Mardia Skewness	4.201	0.379	asymptotic ✓	Normal
2	Mardia Kurtosis	-0.924	0.355	asymptotic ✓	Normal

6ο Ερώτημα:

Αλλάζοντας τον κανόνα ταξινόμησης με τον κανόνα ταξινόμησης του bayes ως προς το 1ο ερώτημα ισχύουν τα ίδια που ίσχυαν και παραπάνω. Στο δεύτερο ερώτημα, υπολογίζουμε της 2 πυκνότητες πιθανότητας, μιας και ακολουθούν πολυμεταβλητή κανονική κατανομή (διμεταβλητή) , με μέσο τον κάθε μέσο του κάθε πληθυσμού και πίνακα συνδυακύμανσης τον pooled covariance matrix. Στην συνέχεια ελέγχουμε, αν η $\frac{f_1}{f_2} > \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$, τότε η παρατήρηση τοποθετείται στον πληθυσμό 1, αντίθετα στον πληθυσμό 2.

Πραγματικό/Predicted	1	2
1	30	0
2	18	27

Όπως παρατηρούμε στον παραπάνω confusion matrix, πλέον εξαιτίας της prior πιθανότητας ο κανόνας ταξινόμησης ευνοεί τον πρώτο πληθυσμό και έτσι δεν υπάρχει δυσταξινόμηση, ως προς τον πρώτο πληθυσμό, αλλά αδικεί τον δεύτερο πληθυσμό, με αποτέλεσμα να υπάρχει μεγαλύτερη δυσταξινόμηση σε σχέση με πριν σε αυτόν. Εν τέλει το ποσοστό λανθάνουσας ταξινόμησης είναι **24%(>14.67%)**, το οποίο είναι σαφώς μεγαλύτερο σε σχέση με την μέθοδο fisher.

Τέλος, εφόσον με την μέθοδο του Fisher, οι νέες παρατηρήσεις ταξινομήθηκαν στον πρώτο πληθυσμό, περιμένω ότι και με την Μπεϋζιανή, η οποία ευνοεί τον πρώτο πληθυσμό έναντι του δευτέρου, να ταξινομήσει τις νέες περιπτώσεις πάλι στον πρώτο πληθυσμό.

Πράγματι, υπολογίζοντας την πυκνότητα πιθανότητας των νέων παρατηρήσεων, ως προς τις κατανομές των πληθυσμών παραπάνω και δεδομένου της prior πιθανότητας, οι 6 παρατηρήσεις κατατάσσονται στον πρώτο πληθυσμό.

Παράρτημα:

1. Όπου αναφέρεται πληθυσμός 1 εννοείται ο πληθυσμός των μη φορέων και πληθυσμός 2, ο πληθυσμός των φορέων

Κώδικας:

```
library(psych)
library(Hmisc)
library(ggplot2)
library(MASS)
library(Morpho)
library(mvtnorm)
library(readxl)
library(klaR)
library(MVN)

## Part A ##

diet=read.csv('///Users//petros//Library//Mobile Documents//com~apple~CloudDocs//MSc Biostatistics//Multivariate Statistics//Assignments//Assignment 1//dietStudy.csv')
deaths=diet$death
diet=diet[,-7]

## 1 ##

## Univariate Descriptive Statistics ##
```

```
summary(diet) ## Means/Medians/Mins/Max

IQR=sapply(diet, IQR, na.rm = TRUE)
Sd=round(sapply(diet, sd, na.rm = TRUE),2)## SD of each variable
Cov= round(Sd^2,2) ## Covariance of each Variable

boxplot(diet,
      main = "Weekly Frequency of Food Consumption",
      ylab = "Times per Week",
      col = "grey",
      las = 2)  ## Boxplots of each Variable

## Multivariate Descriptive Statistics ##

round(cor(diet),2) ## Correlation Matrix

pairs.panels(diet,
         gap = 0,smooth = F,ellipses = F,
         pch = 21) ## Correlation matrix + Histograms + ScatterPlots




## 2 ##

# Testing correlations

KMO(diet)

cortest.bartlett(diet)
FA.PCA <- principal(diet,nfactors = 2,covar = F,rotate = "none")
```

```
# communality
round(rowSums(FA.PCA$loadings^2),3)

## 3 ##

## fit a FA model with varimax rotation ##
FA.PCA1 <- principal(diet,nfactors = 2,covar = F,rotate ="Varimax" )

## fit a FA model with promax rotation ##
FA.PCA2 <- principal(diet,nfactors = 2,covar = F,rotate ="Promax" )

## fit a FA model with oblimin rotation ##
FA.PCA3 <- principal(diet,nfactors = 2,covar = F,rotate ="oblimin" )



plot(FA.PCA$loadings[,1],
   FA.PCA$loadings[,2],
   xlab = "Factor 1",
   ylab = "Factor 2",
   ylim = c(-1,1),
   xlim = c(-1,1),
   main = "No rotation")
abline(h = 0, v = 0)
text(FA.PCA$loadings[,1]-0.08,
   FA.PCA$loadings[,2]+0.08,
   colnames(diet),
   col="blue")
abline(h = 0, v = 0)
```

```
plot(FA.PCA1$loadings[,1],
    FA.PCA1$loadings[,2],
    xlab = "Factor 1",
    ylab = "Factor 2",
    ylim = c(-1,1),
    xlim = c(-1,1),
    main = "Varimax")
abline(h = 0, v = 0)
text(FA.PCA1$loadings[,1]-0.08,
    FA.PCA1$loadings[,2]+0.08,
    colnames(diet),
    col="blue")
abline(h = 0, v = 0)

plot(FA.PCA2$loadings[,1],
    FA.PCA2$loadings[,2],
    xlab = "Factor 1",
    ylab = "Factor 2",
    ylim = c(-1,1),
    xlim = c(-1,1),
    main = "Promax")
abline(h = 0, v = 0)
text(FA.PCA2$loadings[,1]-0.08,
    FA.PCA2$loadings[,2]+0.08,
    colnames(diet),
    col="blue")
abline(h = 0, v = 0)

plot(FA.PCA3$loadings[,1],
    FA.PCA3$loadings[,2],
```

```
    xlab = "Factor 1",
    ylab = "Factor 2",
    ylim = c(-1,1),
    xlim = c(-1,1),
    main = "Oblimin")
abline(h = 0, v = 0)
text(FA.PCA3$loadings[,1]-0.08,
    FA.PCA3$loadings[,2]+0.08,
    colnames(diet),
    col="blue")
abline(h = 0, v = 0)



round(FA.PCA2$Phi,3)

## 4 ##

# Factor scores



FA.PCA2$scores

fa=as.data.frame(FA.PCA2$scores)
fa$deaths=deaths

model1=glm(deaths~RC1+RC2,family = binomial(link ="logit"),data = fa)
summary(model1)



ggplot(fa, aes(x = RC1, y = RC2, color = factor(deaths))) +
  geom_point(alpha = 0.8, size = 0.5) +
```

```
  stat_ellipse(level = 0.95, type = "norm", size = 0.9) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray60") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray60") +
  scale_color_manual(values = c("green", "red"), labels = c("Alive", "Dead")) +
  labs(title = "Factor Scores by Mortality Status",
      x = "Factor 1 (Pescatarian)",
      y = "Factor 2 (Carnivore)",
      color = "Death") +
  theme_minimal(base_size = 14)



par(mfrow=c(1,2))


boxplot(RC1 ~ deaths, data = fa,
        main = "boxplot of factor1~deaths",
        xlab = "Death (0 = No, 1 = Yes)",
        ylab = "Factor1")


boxplot(RC2~ deaths, data = fa,
        main = "boxplot of factor2~deaths",
        xlab = "Death (0 = No, 1 = Yes)",
        ylab = "Factor1")


## 5 ##
eigCor <- eigen(cor(diet))
round(cumsum(eigCor$values/sum(eigCor$values))*100,2)
eigCor$values
diet.fa <- factanal(diet, factors = 2)
diet.fa1 <- factanal(diet, factors = 2,rotation = "varimax")
diet.fa2 <- factanal(diet, factors = 2,rotation = "promax")
```

```
par(mfrow = c(1,1))
hist(diet$meat,
    main = "Histogram of Meat",
    xlab = "Consumption of meat per week")
rowSums(diet.fa$loadings^2)
rowSums(diet.fa1$loadings^2)
rowSums(diet.fa2$loadings^2)



## Part B ##

data1=read_excel("data1.xlsx") ## load Data

## 1 ##

hist(data1$X1,
    main = "Histogram of X1",
    xlab = "X1")
hist(data1$X2,
    main = "Histogram of X2",
    xlab = "X2")



ggplot(data1, aes(x = X1, y = X2, color = factor(group))) +
  geom_point(size = 2.5) +
  labs(x = "X1: log(δραστηριότητα AHF)",
       y = "X2: log(αντιγόνο τύπου-AHF)",
       color = "Ομάδα") +
  scale_color_manual(values = c("1" = "blue", "2" = "red"),
                     labels = c("Μη Φορείς", "Φορείς")) +
```

```
 theme_minimal(base_size = 13)

# Means by group
m1 <- colMeans(data1[data1$group==1,c("X1","X2")])
m2 <- colMeans(data1[data1$group==2,c("X1","X2")])
abs(m2-m1)

## Correlation/Covariance matrix by group
r1 <- cor(data1[data1$group==1,c("X1","X2")])
r2 <- cor(data1[data1$group==2,c("X1","X2")])
s1 <- cov(data1[data1$group==1,c("X1","X2")])
s2 <- cov(data1[data1$group==2,c("X1","X2")])

## Descriptives
sqrt(diag(s1))
sqrt(diag(s2))
max(data1[data1$group==1,"X1"])
max(data1[data1$group==1,"X2"])
max(data1[data1$group==2,"X1"])
max(data1[data1$group==2,"X2"])
min(data1[data1$group==1,"X1"])
min(data1[data1$group==1,"X2"])
min(data1[data1$group==2,"X1"])
min(data1[data1$group==2,"X2"])

## Number of items by group
n1 = sum(data1$group==1)
n2 = sum(data1$group==2)

## Pooled covariance matrix
Sp <- ((n1-1)*s1 + (n2-1)*s2)/(n1+n2-2)
```

```
## 2 ##

## Model Fit
lda1 <- lda(group ~ X1 + X2, data = data1,prior = c(1/2,1/2))

lda1$scaling

## 3 ##

## Model prediction for observations based on the Fisher
pred=predict(lda1)
data1$pred=pred$class

## Confusion Matrix
tab=table(data1$group,data1$pred)

## Percentage of wrongful classification
a=sum(diag(tab))/sum(tab)
1-a

## 4 ##

## Predictions for extra data
data2=read_excel("data 2.xlsx") ## load Data
pred1=predict(lda1,data2)
data2$pred=pred1$class



## 5 ##
```

```
shapiro.test(data1$X1)


shapiro.test(data1$X2)



## multivariate normality test
mvn_result <- mvn(data = data1[, c("X1", "X2")],mvn_test="mardia")
mvn_result$multivariateNormality




b <- -0.5 * t(m1 + m2) %*% lda1$scaling
a <- -lda1$scaling[1] / lda1$scaling[2]  # slope
intercept <- -b / lda1$scaling[2]

# Create decision line as data frame
x_vals <- seq(min(data1$X1), max(data1$X1), length.out = 100)
line_df <- data.frame(
  X1 = x_vals,
  X2 = a * x_vals + intercept
)

# Plot

ggplot() +
  # Original data (with group info)
  geom_point(data = data1,
             aes(x = X1, y = X2,
                 color = factor(group),
                 shape = factor(group)),
```

```
        size = 2.5) +

# New observations
geom_point(data = data2,
        aes(x = X1, y = X2,
          color = "new",
          shape = "new"),
        size = 3) +

# Decision boundary line
geom_line(data = line_df, aes(x = X1, y = X2),
        color = "black", linetype = "dashed", size = 1) +

# Axes and legend titles
labs(x = "X1: log(δραστηριότητα AHF)",
    y = "X2: log(αντιγόνο τύπου-AHF)",
    color = "Κατηγορία",
    shape = "Κατηγορία") +

# Unified legend: color
scale_color_manual(values = c("1" = "blue",
                    "2" = "red",
                    "new" = "green"),
              labels = c("1" = "Μη Φορείς",
                    "2" = "Φορείς",
                    "new" = "Νέες Παρατηρήσεις")) +

# Unified legend: shape
scale_shape_manual(values = c("1" = 16,
                    "2" = 16,
                    "new" = 17),
```

```
            labels = c("1" = "Μη Φορείς",
                       "2" = "Φορείς",
                       "new" = "Νέες Παρατηρήσεις")) +

  theme_minimal(base_size = 13)
```

6

```
##lda2 <- lda(group ~ X1 + X2, data = data1,prior = c(3/4,1/4))
##pred1=predict(lda2)
##data1$pred2=pred1$class

Dens1=dmvnorm(data1[,c("X1","X2")],m1,Sp)
Dens2=dmvnorm(data1[,c("X1","X2")],m2,Sp)

data1$pred1=ifelse(Dens1>Dens2/3,1,2)

## Confusion Matrix
tab1=table(data1$group,data1$pred1)

## Percentage of wrongful classification
a1=sum(diag(tab1))/sum(tab1)
1-a1

## Prediction of new observations
```

```
Dens21=dmvnorm(data2[,c("X1","X2")],m1,Sp)

Dens22=dmvnorm(data2[,c("X1","X2")],m2,Sp)

data2$pred1=ifelse(Dens21>Dens22/3,1,2)
```