Universidad de Granada
Escuela Internacional de Posgrado
Máster en Estadística Aplicada
Materia: Técnicas Estadísticas Multivariantes.
Alumno: Francisco Javier Márquez Rosales

Análisis Discriminante:

Ejercicios:

Diciembre, 2022

1. Realizar un breve resumen de la teoría del Análisis Discriminante con una extensión máxima de 3 páginas.

Análisis Discriminante[1]

El objetivo del análisis discriminante es clasificar los casos de individuos en uno de varios grupos o poblaciones definidas previamente y mutuamente excluyentes, lo que significa que cada individuo debe pertenece a un solo grupo, en función de sus valores para un conjunto de variables predictoras. Se pueden considerar dos fases: análisis y clasificación.

En la fase de análisis, se desarrolla una regla de clasificación utilizando casos para los que se conoce la pertenencia a grupos. En la fase de clasificación, la regla se utiliza para clasificar los casos para los que no se conoce la pertenencia a grupos. La variable de agrupación debe ser categórica y las variables (predictoras) independientes deben ser de intervalo o dicotómicas, ya que se utilizarán en una ecuación de tipo regresión. Por esto, estas variables deben cumplir con los mismos supuestos que se aplican para esta técnica.

El punto de partida para realizar análisis discriminante es disponer de una matriz de datos de N individuos, con p variables cuantitativas independientes (explicativas o discriminantes), que definen el perfil de cada individuo y una variable adicional, cualitativa, que es la variable dependiente, la que contiene la información del grupo de pertenencia para cada uno de los individuos.

Clasificación a partir de una variable discriminante:

Supongamos que disponemos de una variable discriminante X para clasificar un nuevo individuo entre dos grupos posibles G_I y G_{II}. Nuestro objetivo es encontrar

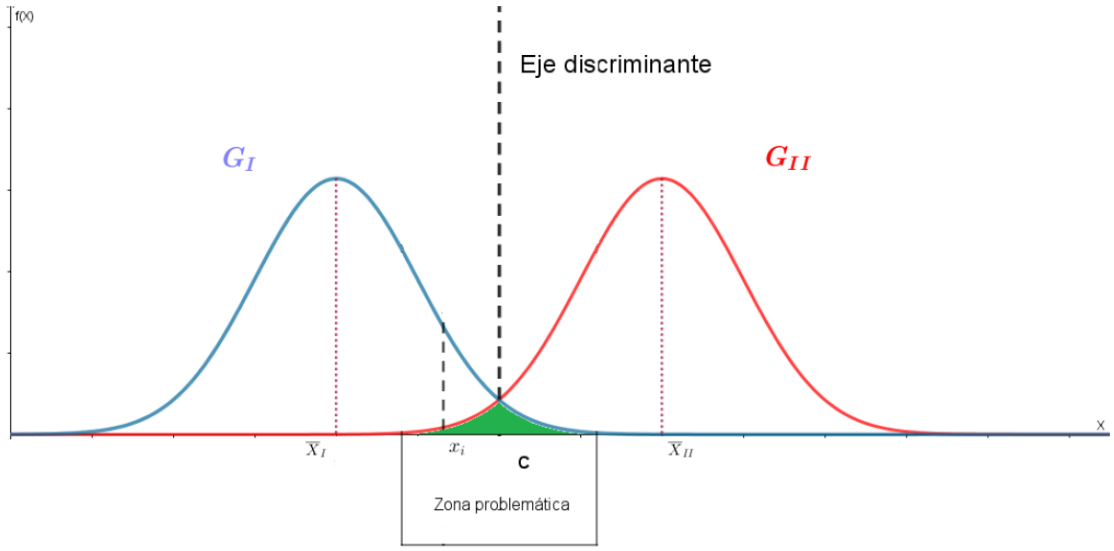
[1] Este resumen está basado principalmente en el documento: Tema 7. Análisis Discriminante. Proporcionado como documento Guia en la Materia Análisis Multivariante. Postgrado en Estadística Aplicada. Universidad de Granada.

una función lineal de la variable discriminante X que nos permita clasificar cada observación en uno de los grupos, minimizando el error de clasificación.

Esta situación gráficamente aparece en la Imagen 1.

Se asume que la distribución de X en cada uno de los grupos solo se diferencia en su posición, ya que tiene la misma forma y dispersión para cada grupo.

Imagen 1. Clasificación a partir de una variable discriminante



Clasificación a partir de dos variables discriminantes:

Queremos encontrar una función lineal de las variables discriminantes X1 y X2 que permita clasificar cada observación en G_I o en G_{II}, minimizando el error de clasificación.

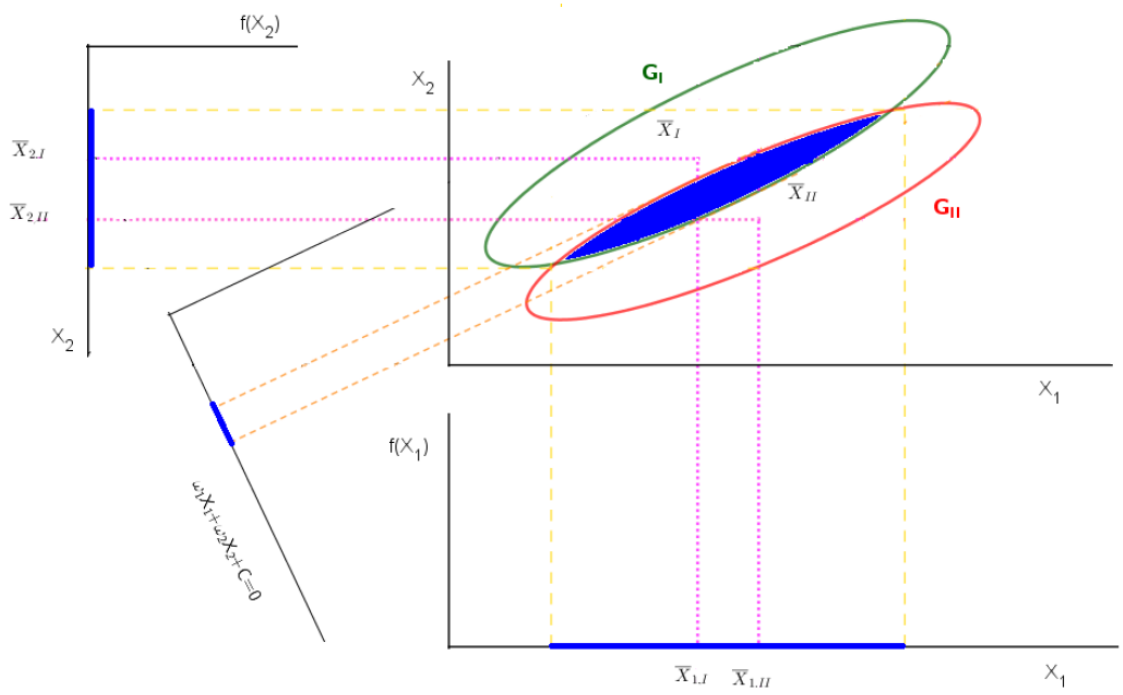
Consideramos en primer lugar la variable X1 y proyectamos sobre el eje correspondiente los datos observados, al igual que lo hicimos en el caso de una variable discriminante. Al hacer esto, se crea una zona problemática debido a la superposición de las distribuciones normales de los dos grupos. Luego, se hace lo mismo con X2, creando otra nueva zona problemática. Por tanto, al proyectar en cada uno de los ejes (variables) tenemos dos zonas problemáticas

Luego, se crea una función que maximiza la separación entre los grupos, maximizando la distancia entre sus medias, y minimizando la dispersión dentro de los grupos. Con la forma

$$D=w_1X_1+w_2X_2$$

La representación gráfica de este problema puede verse en la Imagen 2.

Imagen 2. Función lineal que minimiza la zona problemática con dos variables discriminantes.



Clasificación a partir de p variables discriminantes:

El objetivo es encontrar una función discriminante capaz de minimizar la variabilidad dentro de los grupos y maximizar la variabilidad entre los grupos y que sea combinación lineal de las p variables de las que se dispone:

$$D=w_1X_1+w_2X_{2+}\ \ldots\ +w_pX_p$$

el caso de tener N observaciones. Para cada observación $i = 1, \ldots, N$ la función discriminante tiene la siguiente forma:

$$D=w_1X_{1i}+w_2X_{2i+}\ \ldots\ +w_pX_{pi}$$

2. EJERCICIO. Se desea estudiar la ansiedad de un grupo de ex-fumadores. Para ello se clasifica la ansiedad en tres grupos (1, 2 y 3) según la intensidad con que se manifiesten los síntomas de esta. Se obtiene una muestra para cada grupo y se les miden a todos los individuos 3 variables X1, X2 y X3 relacionadas con sus esquemas de comportamiento. Se desea obtener una función discriminante lineal para poder clasificar en un grupo u otro a un individuo, en base a las variables X1, X2 y X3.

Realizar un análisis discriminante completo y predecir a qué grupo de ansiedad pertenecerá con mayor probabilidad un individuo con valores X1=7.5, X2=9 y X3=2.

Solución:

En primer lugar, hacemos la lectura de los datos y resumimos los mismos para validar la correcta lectura.

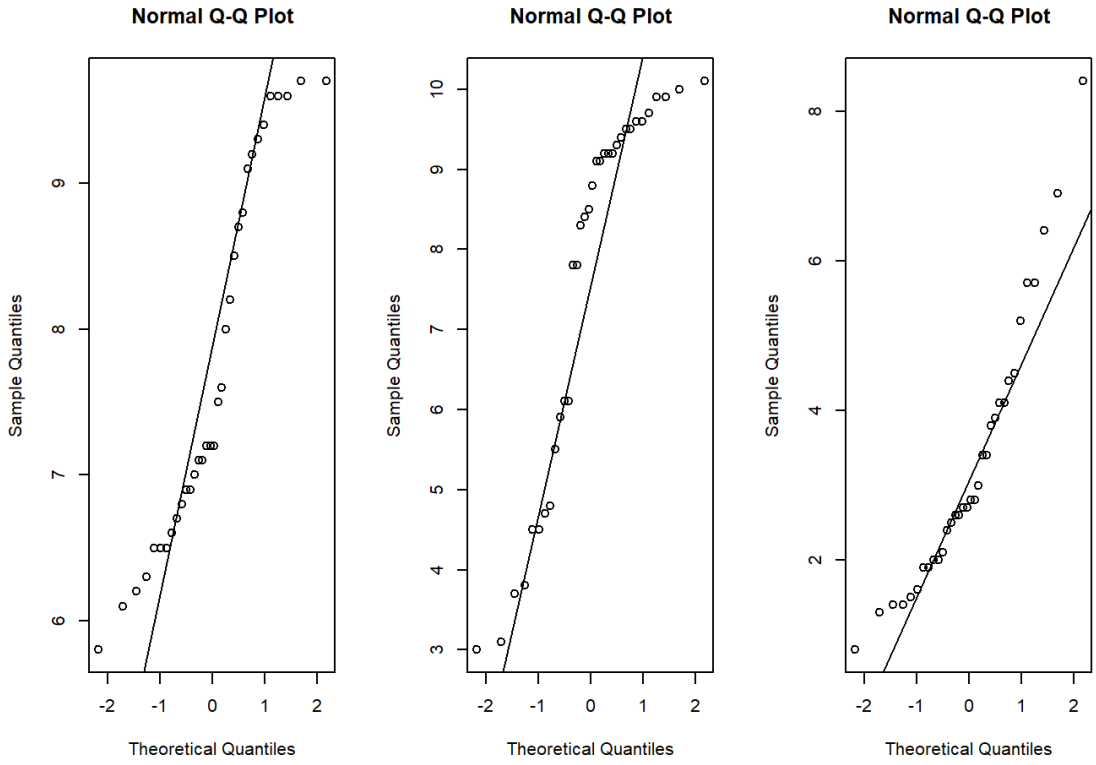
```
datos<-read.table("discriminante2.txt", header=TRUE)
attach(datos)
str(datos)
## 'data.frame':    34 obs. of  4 variables:
##  $ ansiedad: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ X1      : num  6.5 6.2 5.8 6.5 6.5 6.9 7.2 6.9 6.1 6.3 ...
##  $ X2      : num  9.5 9.9 9.6 9.6 9.2 9.1 10 9.9 9.5 9.4 ...
##  $ X3      : num  4.4 6.4 3 4.1 0.8 5.7 2 3.9 1.9 5.7 ...
```

Ahora, hacemos inicialmente la comprobación de las hipótesis sobre los datos, que validarían la aplicación de la técnica:

Hipótesis: Normalidad

```
library(mvnormtest)
shapiro.test(t(datos[,2:4]))
##  Shapiro-Wilk normality test
## data:  t(datos[, 2:4])
## W = 0.92656, p-value = 2.682e-05
par(mfrow = c(1,3))
qqnorm(datos$X1)
qqline(datos$X1)
```

```
qqnorm(datos$X2)
qqline(datos$X2)
qqnorm(datos$X3)
qqline(datos$X3)
```




```
par(mfrow = c(1,1))
```

A partir del resultado del Test Shapiro, p=2.682e-05, se rechaza la hipótesis nula, lo que sugiere que las variables X1, X2 y X3 no presentan un comportamiento normal.

Hipótesis: Igualdad de varianzas-covarianzas

```
homocedasticidad_tratamiento=bartlett.test(list(datos$X1, datos$X2, datos$X3))
print(homocedasticidad_tratamiento)
```

```
##  Bartlett test of homogeneity of variances
## data:  list(datos$X1, datos$X2, datos$X3)
## Bartlett's K-squared = 11.978, df = 2, p-value = 0.002506
```

El resultado del test de Barlett sugiere el rechazo de la hipótesis nula, en otras palabras, indica que las variables no representan igualdad de varianzas.

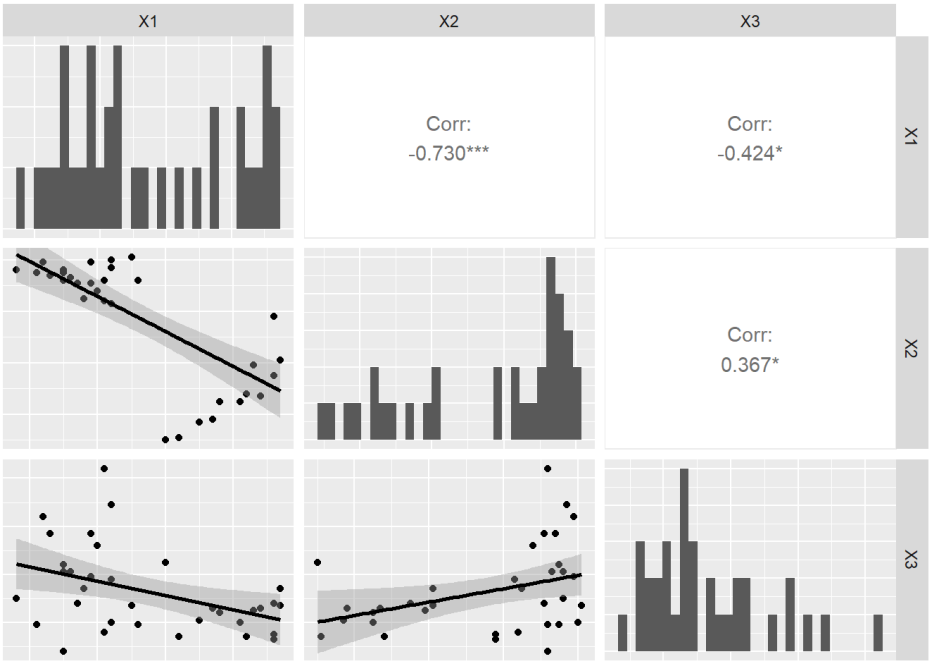
Hipótesis: No multicolinealidad

```
cor(datos[c("X1","X2","X3")], use="complete")
```

```
##             X1         X2         X3
## X1  1.0000000 -0.7300793 -0.4236522
## X2 -0.7300793  1.0000000  0.3670388
## X3 -0.4236522  0.3670388  1.0000000
```

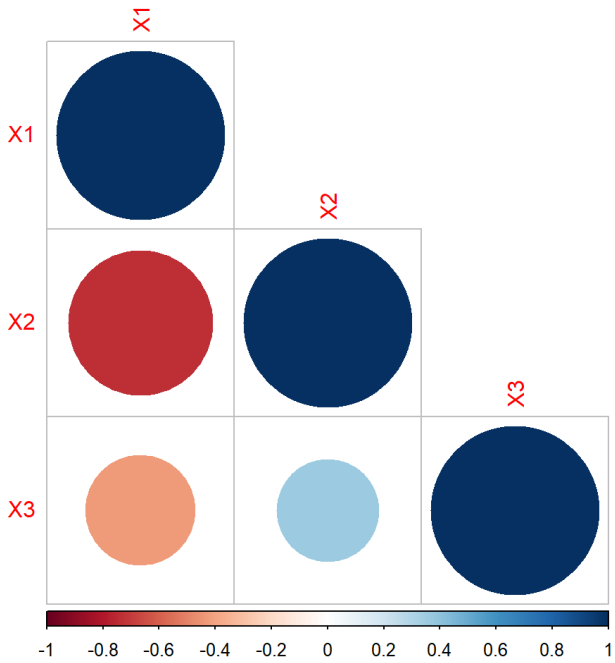
```
ggpairs(datos[,2:4], lower = list(continuous = "smooth"),
diag = list(continuous = "barDiag"), axisLabels = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
corrplot(cor(datos[c("X1","X2","X3")], use="complete"),sig.level=0.05,typ="lower")
```

Al ver la matriz de covarianzas así como el grafico (a continuación), vemos que efectivamente entre las variables X1 y X2 podemos tener colinealidad por valores que indican alto grado de relación bivariante. Para el resto de las combinaciones no se presenta este comportamiento.



Ahora iniciamos el análisis Discriminante:

Consideramos que las probabilidades de pertenencia a cada grupo son iguales. Usaremos, además, el método de los momentos para la estimación

```
discrimi<-lda(ansiedad~X1+X2+X3,prior=c(0.33,0.33,0.34), method="moment", tol=0.001)
discrimi
```

```
## Call:
## lda(ansiedad ~ X1 + X2 + X3, prior = c(0.33, 0.33, 0.34), method = "moment",
##     tol = 0.001)
##
## Prior probabilities of groups:
##    1    2    3
## 0.33 0.33 0.34
##
## Group means:
##     X1       X2       X3
## 1 6.49 9.570000 3.790000
## 2 7.08 9.060000 4.080000
## 3 9.10 5.092857 2.371429
##
## Coefficients of linear discriminants:
##             LD1        LD2
## X1  2.763113352 -0.8926737
```

```
## X2 -1.180044092 -0.3316161
## X3 -0.009738474 -0.4716923
##
## Proportion of trace:
##    LD1    LD2
## 0.9986 0.0014
```

Del resultado obtenemos lo siguiente:

La tabla con las probabilidades a priori de pertenencia a cada grupo. Los centroides de los grupos. Los coeficientes de las funciones discriminantes. En nuestro caso dos en función de las 3 variables.

Resultando las siguientes funciones:

D1= +2.763X1-1.180X2-0.009X3

D2= -0.892X1-0.331X2-0.471X3

Ahora, obtenemos la proporción de varianza explicada por cada eje

```
discrimi$svd
```

```
## [1] 22.5655057  0.8395045
```

Comprobamos que la segunda función es mucho más discriminante que la primera.

Ahora realizamos la predicción sobre el individuo.

```
Datos2<-rbind(c(7.5, 9, 2))
Datos21<-data.frame(Datos2)
discrimi2<-predict(discrimi,newdata=Datos21,prior=discrimi$prior,2)
discrimi2
```

```
## $class
## [1] 2
## Levels: 1 2 3
##
## $posterior
##             1         2            3
## 1 0.006135785 0.9938642 5.206912e-18
##
## $x
```

```
##         LD1       LD2
## 1 -1.507827 0.3547427
```

El resultado nos indica que el individuo tiene probabilidad de 0.993 de pertenecer al grupo 2, lo que confirmaría su pertenencia al grupo N2 de acuerdo a su ansiedad de fumar.

Obtenemos ahora la matriz de confusión

```
table(predict(discrimi)$class, ansiedad)
##    ansiedad
##       1  2  3
##   1   8  2  0
##   2   2  8  0
##   3   0  0 14
```

El resultado nos indica que, según el modelo, hay 2 individuos mal clasificados como del grupo 1 y dos individuos mal clasificados como del grupo 2.

A continuación, obtenemos algunos gráficos asociados al análisis.

Gráfico de las puntuaciones discriminantes

```
plot(discrimi)
```

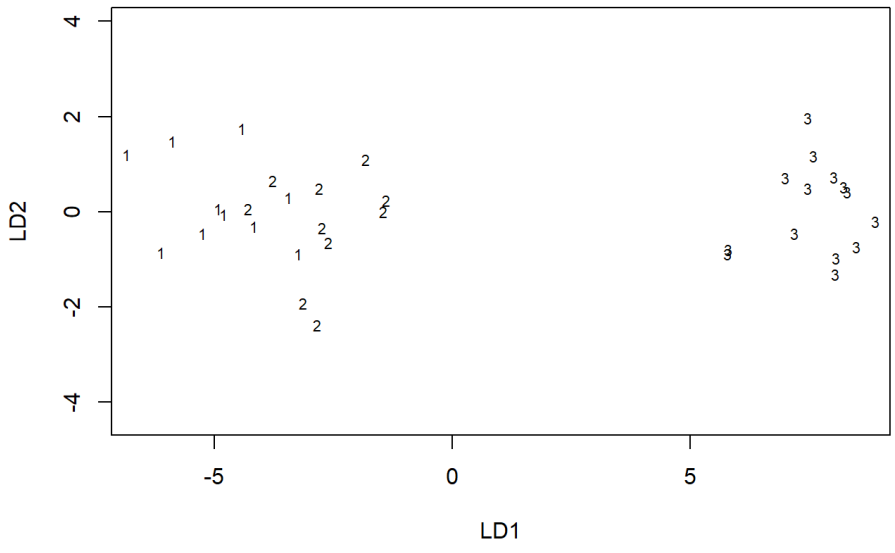
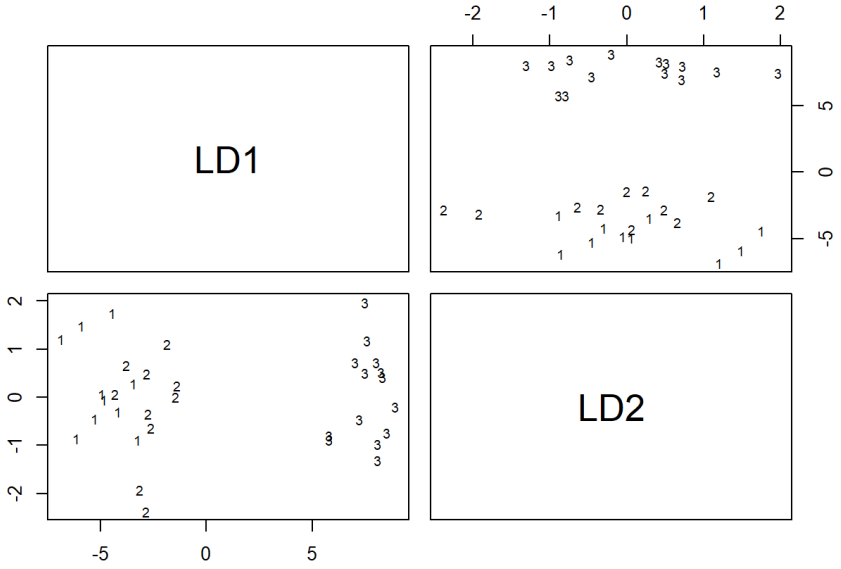


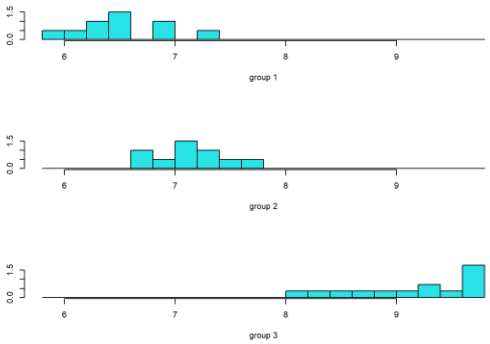
Gráfico de las puntuaciones discriminantes de los datos

```
pairs(discrimi)
```

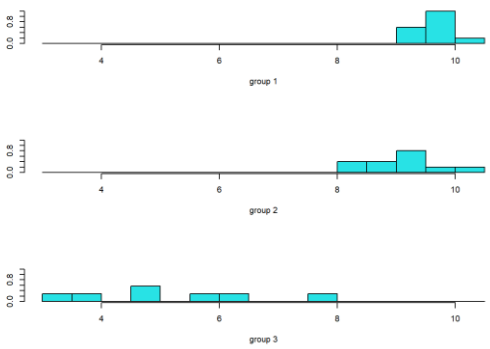


Finalmente, obtenemos el histograma de las variables dependientes frente a la variable de agrupación.

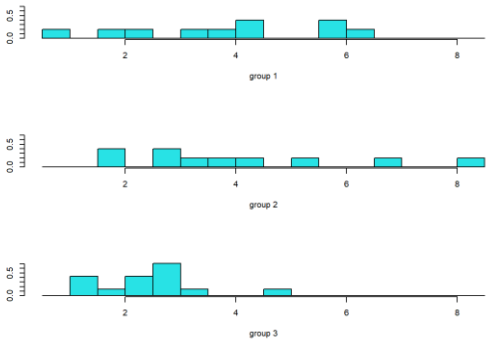
```
ldahist(datos$X1,datos$ansiedad)
```




```
ldahist(datos$X2,datos$ansiedad)
```




```
ldahist(datos$X3,datos$ansiedad)
```



Validez del modelo

Al principio del análisis vimos como no se cumplen los supuestos de normalidad y variabilidad constante para el conjunto estudiado. Esto no nos permite declarar como valido el modelo planteado. Una opción que podemos tomar para poder completar de forma adecuada el análisis seria hacer exploraciones para encontrar una transformación adecuada de los datos que nos permitan validar los supuestos.