Métodos Multivariados II

Centro de Investigación en Matemáticas A.C.

www.cimat.mx

Apéndice E: Análisis Discriminante en R

El Análisis Discriminante es una técnica multivariada que se emplea cuando se necesita asignar una o varias observaciones a grupos o clústers conocidos previamente (*a priori*) a partir de los valores de un conjunto de variables medidas sobre las observaciones a los que se pretende clasificar.

La importancia de este tipo de problemas multivaridados radica en la necesidad de las personas para establecer estrategias que le permitan clasificar correctamente objetos o individuos buscando disminuir las consecuencias negativas en las que se pudiera incurrir por una mala clasificación. Por ejemplo, para un anestesiólogo es importante no cometer una equivocación al clasificar un paciente como apto para que se le suministre la anestesia, puesto que sería una equivocación sumamente grave que podría causar la muerte el clasificar un paciente como seguro para un anestésico, cuando en realidad es inseguro para él.

El Análisis Discriminante es equivalente al Análisis de Regresión solo que la pertenencia de un individuo a uno u otro grupo se introduce mediante una variable dependiente categórica que tiene como valores la etiqueta de cada uno de los grupos. Las variables independientes o también denominadas variables clasificadoras, predictoras o explicativas son continuas y permiten determinar a qué grupos pertenecen los objetos.

El objetivo básico de esta técnica es sintetizar la información de las variables predictoras en relaciones lineales denominadas funciones discriminates que mejor discriminen a los grupos. Las funciones discriminantes describen e interpretan las diferencias que existen entre dos o más poblaciones (clústers). Además, se trata de definir una regla de decisión que asigne un objeto nuevo, no clasificado previamente, a uno de los grupos prefijados en función de las mediciones efectuadas en las variables predictoras. La "mejor" regla será aquella que tenga la menor tasa de clasificaciones erróneas. En términos estadísticos equivale a decir que es preferible aquella regla que tenga la más baja tasa de error de todas las futuras asignaciones de individuos.

En resumen, el Análisis Discriminante el análisis discriminante busca, por un lado, establecer una función discriminante que maximice la separación (diferencias) existentes entre las agrupaciones de individuos consideradas, y por el otro establecer reglas de clasificación que minimicen la tasa de ocurrencia de clasificaciones erróneas de casos futuros.

Algunos ejemplos de problemas donde ha sido aplicado exitosamente el Análisis Discriminante son:

- asignar un cráneo a una de dos posibles especies animales,
- asignar un texto escrito a uno de entre dos posibles autores,
- decidir que una declaración de impuestos es potencialmente defraudadora o no,
- determinar que una empresa está en riesgo de quiebra o no,
- decidir que un nuevo método de fabricación es eficaz o no.

Existen varios enfoques posibles para establecer las funciones discriminantes para abordar el problema de clasificación. Uno de ellos y el más empleado es el Análisis Discriminante Lineal clásico desarrollado por Fisher. Este procedimiento está basado en la normalidad multivariante de las variables y es óptimo bajo dicho supuesto.

A continuación, se describe el procedimiento a seguir para ejecutar el Análisis Discriminante Lineal propuesto por Fisher en el paquete estadístico R.

E.1. Discriminación para dos poblaciones

Siguiendo con el ejemplo introductorio de la Unidad 2, se desea crear la "regla" que permita separar y clasificar a los estudiantes de nuevo ingreso en dos grupos ($\mathbf{X}_1$): los que logran titularse (π_1) y aquellos que no terminan sus estudios (π_2). Las variables predictoras que se utilizarán en el análisis son las puntuaciones obtenidas en dos pruebas de ingreso:

$\mathbf{X_2}$: Evaluación del conocimiento general del estudiante. El valor máximo de esta prueba es de 700 puntos.
$\mathbf{X_3}$: Evaluación sobre el nivel de desarrollo de las aptitudes intelectuales y características emocionales. Los valores van en un rango de 1 a 4.

El archivo *IngresoUniversidad.xlsx* contiene los registros de las puntuaciones obtenidas por 85 alumnos de generaciones anteriores y el resultado de su paso por la universidad. Los estudiantes están clasificados en tres grupos: los que lograron titularse, los que no lo terminaron los estudios y aquellos que están en el límite de alcanzar la titulación. Para este primer ejemplo solo se consideran los dos primeros grupos.

E.1.1. Verificación de los supuestos

Se da inicio con la importación de la base de datos al programa R. Recuerde colocar la ruta correcta (sección en color verde) en su computadora donde tiene guardado el archivo antes mencionado.

```
library(readr)
IngresoUniversidad <- read_excel("G:/Mi unidad/CIMAT/EME/MULTIVARIADOS
II/PRED/MATERIAL/DISCRIMINANTE/ IngresoUniversidad.xlsx")
View(IngresoUniversidad)
```

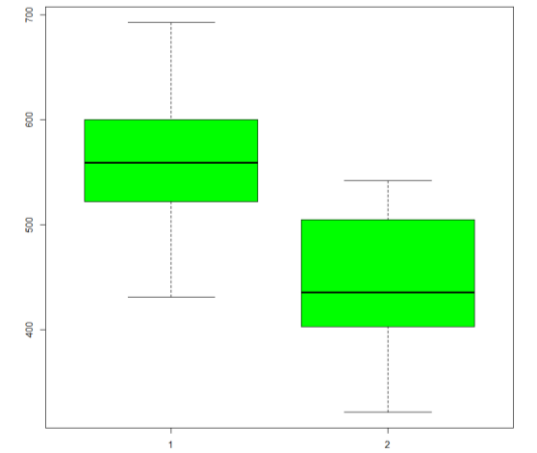
La columna nombrada "Resultado" indica el grupo al que pertenece cada una de las observaciones. El grupo 1 son los estudiantes que lograron titularse, el grupo 2 contiene los resultados de aquellos que no terminaron la universidad, y el tercer grupo engloba a los casos que están en el límite de alcanzar la titulación. Para este ejemplo, como se mencionó únicamente se estudiarán a los grupos 1 y 2, los cuales suman 59 expedientes. Se verifica la cantidad de expedientes que contiene cada grupo.

```
table(IngresoUniversidad$Resultado[1:59])

1  2
31 28
```

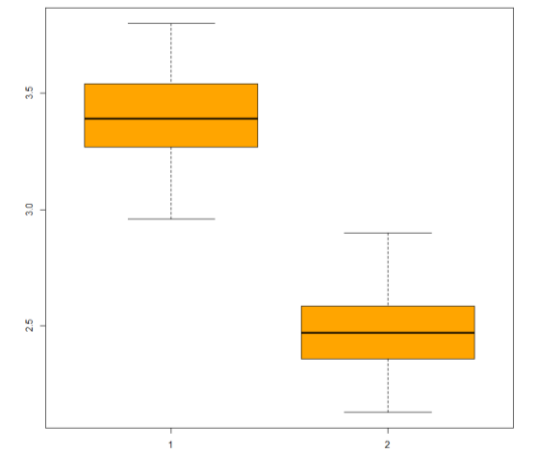
El grupo de los titulados consta de 31 expedientes, mientras 28 estudiantes de la base de datos no concluyeron sus estudios. Como se menciona en el material de la unidad, el primer paso es verificar todos los supuestos antes de ejecutar el análisis. Se da inicio con un análisis descriptivo de los datos para verificar si es posible distinguir la separación de los grupos bajo estudio.

```
summary(IngresoUniversidad[IngresoUniversidad$Resultado == 1,-c(1,2)])

  Conocimiento     Aptitudes
 Min.   :431.0   Min.   :2.960
 1st Qu.:522.0   1st Qu.:3.270
 Median :559.0   Median :3.390
 Mean   :561.2   Mean   :3.398
 3rd Qu.:600.5   3rd Qu.:3.540
 Max.   :693.0   Max.   :3.800


summary(IngresoUniversidad[IngresoUniversidad$Resultado == 2,-c(1,2)])

  Conocimiento     Aptitudes
 Min.   :321.0   Min.   :2.130
 1st Qu.:404.2   1st Qu.:2.360
 Median :435.5   Median :2.470
 Mean   :447.1   Mean   :2.482
 3rd Qu.:504.2   3rd Qu.:2.578
 Max.   :542.0   Max.   :2.900
```

```
boxplot(IngresoUniversidad$Conocimiento[IngresoUniversidad$Resultado ==
1],IngresoUniversidad$Conocimiento[IngresoUniversidad$Resultado == 2],
col = "green")
```

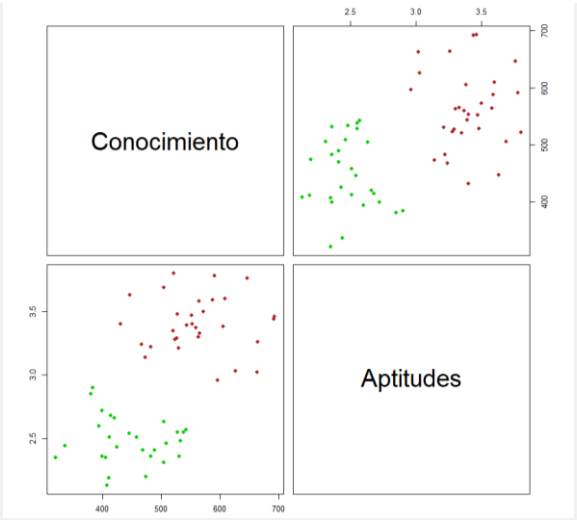



```
boxplot(IngresoUniversidad$Aptitudes[IngresoUniversidad$Resultado ==
1], IngresoUniversidad$Aptitudes[IngresoUniversidad$Resultado == 2],col
= "orange")
```



Se aprecia en los gráficos boxplot que para los puntajes de la prueba de aptitudes intelectuales y emocionales existe una distinción muy clara entre los dos grupos, por el contrario, al comparar los puntajes de la prueba de conocimiento hay un traslape entre los dos grupos. En ocasiones el poder de separación de las variables predictoras individuales dificulta el encontrar esa regla para la mejor separación, pero al conjuntar dos o más variables la separación puede ser más evidente. Se realiza el gráfico de dispersión para validar esta conclusión.

```
IngresoUniversidad$Resultado <- factor(IngresoUniversidad$Resultado)
pairs(x = IngresoUniversidad[1:59, c("Conocimiento","Aptitudes")],
      col = c("firebrick", "green3")[IngresoUniversidad$Resultado],
      cex.labels = 3, pch = 19)
```

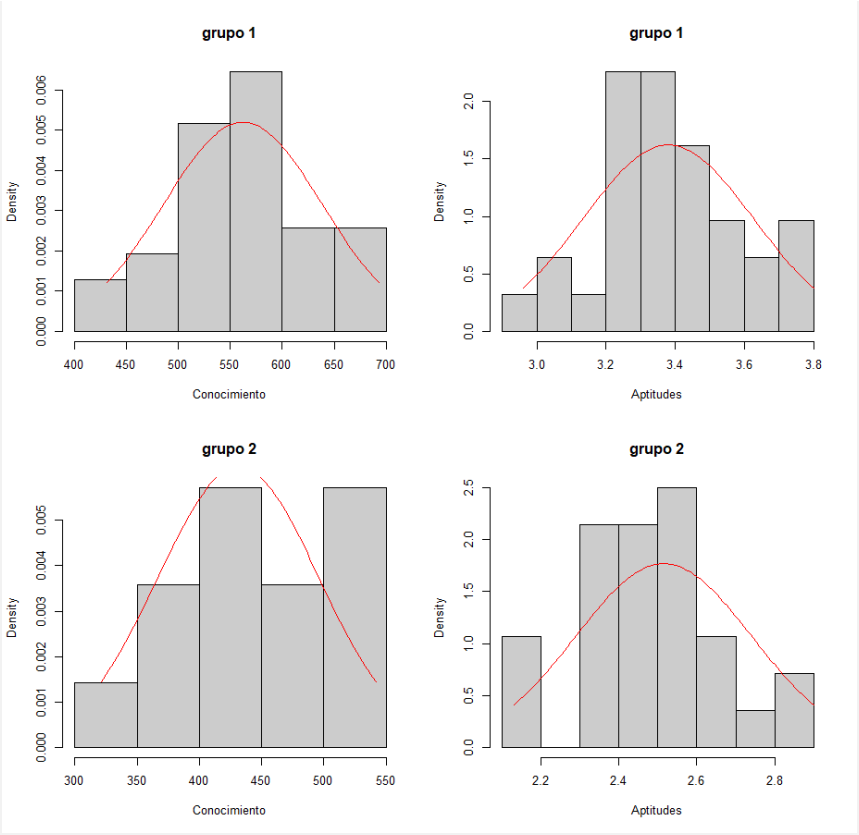


El siguiente paso es comprobar la normalidad univarinate de las variables predictoras, para ello se puede realizar los histogramas correspondientes o utilizar la prueba de hipótesis para normalidad de Shapiro – Wilk la cual evalúa que:

H$_0$: la muestra de valores x_1, x_2, …, x_n proviene de una población normalmente distribuida.
H$_1$: la muestra de valores x_1, x_2, …, x_n NO proviene de una población normalmente distribuida.

Para el caso del Análisis Discriminante lo que se estaría buscando es que la hipótesis nula no fuese rechazada, es decir, que el p–valor sea mayor al nivel de significancia seleccionado.

```r
# Representación mediante Histograma de cada variable para cada grupo
par(mfcol = c(2, 2))
for (k in 3:4) {
  j0 <- names(IngresoUniversidad)[k]
  for (i in 1:2) {
    x <- IngresoUniversidad[IngresoUniversidad$Resultado == i, j0]
    hist(as.matrix(x), freq = F, col = grey(0.8), main = paste("grupo",
         i),xlab = j0)
    curve(dnorm(x,mean(as.matrix(x)),sd(as.matrix(x))),
          min(as.matrix(x)), max(as.matrix(x)),add = T,col="red")
  }
}
par(mfcol = c(1, 1))
```




```
# Contraste de normalidad Shapiro-Wilk para cada variable en cada grupo
for (k in 3:4) {
  j0 <- names(IngresoUniversidad)[k]
  for (i in 1:2) {
    x <- IngresoUniversidad[IngresoUniversidad$Resultado == i, j0]
    x.test <- shapiro.test(as.matrix(x))
    x.test$data.name <- paste("grupo", i,",",j0)
    print(x.test)
  }
}
```

```

    Shapiro-Wilk normality test

data:  grupo 1 , Conocimiento
W = 0.9775, p-value = 0.7403


    Shapiro-Wilk normality test

data:  grupo 2 , Conocimiento
W = 0.94627, p-value = 0.1595


    Shapiro-Wilk normality test

data:  grupo 1 , Aptitudes
W = 0.97932, p-value = 0.7935


    Shapiro-Wilk normality test

data:  grupo 2 , Aptitudes
W = 0.97996, p-value = 0.8496
```

De los resultados se desprende que las variables predictoras en cada uno de los grupos cumplen con el supuesto de normalidad univariada. El test de normalidad concluye que para las evaluaciones realizadas ninguno de los cuatros casos se rechaza la hipótesis nula al considerar un nivel de significancia de 0.05.

Corresponde ahora contrastar la homogeneidad de la varianza para lo cual se utilizará en este caso la prueba de M-Box propuesta por el matemático Box en 1949. Esta prueba es utilizada en investigaciones multivariantes para contrastar la igualdad de matrices entre grupos. El test M-Box es muy sensible a violaciones de la normalidad multivariante, por lo que debe ser contrastado con anterioridad. La prueba evalúa las siguientes hipótesis:

$\mathbf{H_0}$: $\sum_1 = \sum_2 = \ldots = \sum_k$ (la matriz de covarianza de las k poblaciones o grupos son todas iguales)
$\mathbf{H_1}$: la matriz de covarianza de las k poblaciones o grupos NO son todas iguales

Al igual que en la prueba anterior lo que se busca es NO rechazar estadísticamente la hipótesis nula.

```
# Prueba de M-Box
install.packages("biotools")
library("biotools")
boxM(IngresoUniversidad[1:59,-c(1:2)],IngresoUniversidad
$Resultado[1:59])

    Box's M-test for Homogeneity of Covariance Matrices

data:  IngresoUniversidad[1:59, -c(1:2)]
Chi-Sq (approx.) = 1.0439, df = 3, p-value = 0.7906
```

Después de comparar el *p*–valor arrojado por la prueba con el nivel de significancia de 0.05 se concluye que no hay suficiente evidencia estadística para rechazar la afirmación de que las matrices de covarianza de los dos grupos son iguales.

A continuación se verifica la multicolinealidad de las variables en cada grupo para lo cual se emplea la matriz de correlación. En caso de encontrar valores superiores a 0.8 en valor absoluto se deben excluir del análisis.

```
# Matriz de correlación por grupo
round(cor(IngresoUniversidad[IngresoUniversidad$Resultado == 1,
-c(1:2)]),2)

             Conocimiento Aptitudes
Conocimiento         1.00     -0.04
Aptitudes           -0.04      1.00

round(cor(IngresoUniversidad[IngresoUniversidad$Resultado == 2,
-c(1:2)]),2)

             Conocimiento Aptitudes
Conocimiento          1.0      -0.1
Aptitudes            -0.1       1.0
```

No se aprecia problemas de multicolinealidad en alguno de los dos grupos. Por lo tanto se puede concluir que es factible realizar el Análisis Discriminante para este caso.

E.1.2. Estimación del modelo discriminante

Se procede a calcular la función discriminante utilizando la aproximación de Fisher. Para ello, se utiliza el comando `lda()`. Se debe especificar la relación entre la variable categórica dependiente y las variables predictoras continuas seleccionadas.

```
install.packages("MASS")
library("MASS")
# Se crea la relación entre la variable categorica y las variables
# predictoras
IngresoUniversidad.lda <-
   lda(formula = Resultado ~ Conocimiento + Aptitudes,
   data = IngresoUniversidad[1:59,-1])

# Se despliegan los resultados del modelo creado
IngresoUniversidad.lda

Call:
lda(Resultado ~ Conocimiento + Aptitudes, data = IngresoUniversidad
[1:59,-1])

Prior probabilities of groups:
        1         2
0.5254237 0.4745763

Group means:
  Conocimiento Aptitudes
1     561.2258  3.398387
2     447.0714  2.482500

Coefficients of linear discriminants:
                      LD1
Conocimiento -0.006275466
Aptitudes    -4.655542832
```

Los resultados del análisis indican que la prioridad a priori de pertenecer a alguno de los grupos. El comando calcula los valores considerando el tamaño de cada grupo con respecto al total de casos de prueba utilizados para ejecutar el análisis. En el ejemplo, la probabilidad de pertenecer al grupo 1 es de 52.54% (31/59) y para el grupo 2 la probabilidad es de 47.46% (28/59). En caso de que se conozca la verdadera probabilidad ya sea porque los resultados de otros estudios o por experiencia se puede indicar introduciendo el argumento `prior` en el comando `lda()`.

Se despliega el valor promedio de cada variable predictora en cada uno de los grupos considerados, así como los coeficientes de la función de discriminante lineal de Fisher. La ecuación quedaría expresada de la siguiente manera:

$$-0.006 * \text{Conocimiento} - 4.656 * \text{Aptitud}$$

En donde, *Conocimiento* y *Aptitud* son los valores de los puntajes obtenidos en dichas pruebas. Por conveniencia, los valores de las funciones discriminantes estimados está reescalado considerando una media igual a 0.

E.1.3. Valorar la exactitud de la población

Siguiendo con las etapas del análisis corresponde analizar la precisión del modelo con respecto a la tasa de casos mal clasificados. Para ello se toma la misma muestra utilizada para generar el modelo discriminante.

```
# Predictor de valores utilizando el modelo discriminante
IngresoUniversidad.lda.values <- predict(IngresoUniversidad.lda)

# Tabla de clasificación de datos reales contra los predichos
table(IngresoUniversidad$Resultado[1:59],IngresoUniversidad.lda.
values$class, dnn = c("Clase real", "Clase predicha"))

           Clase predicha
Clase real  1  2
         1 31  0
         2  0 28


# Calculo de la tasa de error de clasificación
tasa_error <- mean(IngresoUniversidad$Resultado[1:59] !=
IngresoUniversidad.lda.values$class) * 100
paste("tasa_error (mala clasificación)=", tasa_error, "%")


[1] "tasa_error (mala clasificación)= 0 %"
```

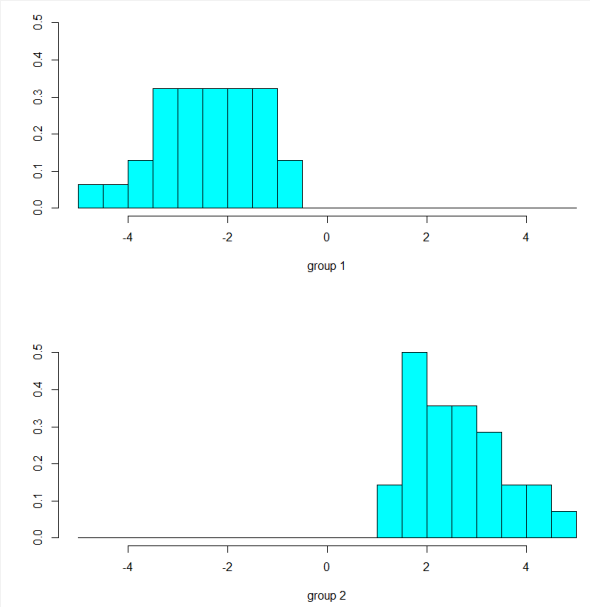
La evaluación arroja que la regla creada clasifica correctamente a todos los casos de la muestra, es decir la tasa de error de 0%. Es importante que para tener una confirmación del modelo obtenido para lo cual se debe tener otro grupo de casos correctamente clasificados y comprobar la tasa de error. Se puede separar desde el inicio del estudio a la muestra en el grupo de entrenamiento, sobre el cual se aplicará la metodología descrita, y el grupo de prueba, que permitirá su validación.

Para conocer los valores obtenidos de cada uno de los individuos en al evaluarlos en la función discriminante se precede de la forma siguiente:

```
IngresoUniversidad.lda.values$x

           LD1                        30 -4.5790453
1  -0.5408377                         31 -1.0348546
2  -0.6069531                         32  2.3558102
3  -1.0358757                         33  2.9997046
4  -1.6441597                         34  3.7629817
5  -3.3683166                         35  2.6603933
6  -3.4773293                         36  1.6136991
7  -1.0549897                         37  3.4913819
8  -1.2406265                         38  2.7088423
9  -2.7250070                         39  2.4201708
10 -3.4236260                         40  3.4887548
11 -1.9166319                         41  2.9678911
12 -2.3194315                         42  1.9603071
13 -2.9042197                         43  1.9048491
14 -4.3270055                         44  2.0891772
15 -3.3779430                         45  2.3328992
16 -2.5349883                         46  1.5728343
17 -2.6390441                         47  3.5116657
18 -1.8795640                         48  4.5030504
19 -2.2101214                         49  2.8166950
20 -1.5725024                         50  1.7319119
21 -1.2905427                         51  3.0563325
22 -3.2264594                         52  2.6911857
23 -2.0688491                         53  4.2048915
24 -1.5538247                         54  4.0247965
25 -2.5526449                         55  2.4028018
26 -2.3642323                         56  1.7946665
27 -3.6019662                         57  1.8127593
28 -2.2174180                         58  1.3204972
29 -3.9808338                         59  1.0688936
```

Otra manera de desplegar los resultados del análisis discriminante es hacer un histograma con los valores obtenidos después de evaluar la función discriminante para las observaciones de la muestra de los diferentes grupos.

```
# Histograma de los valores predichos
ldahist(IngresoUniversidad.lda.values$x[,1], g = IngresoUniversidad$
Resultado[1:59])"
```



En el histograma se aprecia que la única función discriminante estimada separa claramente a los dos grupos. Los valore de la función para el primer grupo van de -5 a -1 mientras que para el grupo 2 oscila entre 1 y 5, por lo que no existe traslapes.

E.1.4. Clasificación de nuevos casos

Como se mencionó en la introducción otra de las características de este método multivariante es el poder utilizar la función discriminante (regla) para clasificar nuevas observaciones. Como ejemplo, supóngase que hay tres nuevos expedientes que necesitan ser clasificados (ver archivo *IngresoUniversidad_NvoCasos.xlsx*)

```
# Importar la base donde están los nuevos casos
IngresoUniversidad_NvoCasos <- read_excel("G:/Mi unidad/CIMAT/EME/
MULTIVARIADOS II/PRED/MATERIAL/DISCRIMANANTE/IngresoUniversidad_
NvoCasos.xlsx")
View(IngresoUniversidad_NvoCasos)

predict(object = IngresoUniversidad.lda, newdata = IngresoUniversidad
        _NvoCasos)

$`class`
[1] 1 2 1
Levels: 1 2

$posterior
            1            2
1 0.775688056 2.243119e-01
2 0.001061859 9.989381e-01
3 0.999999905 9.517704e-08

$x
         LD1
1 -0.1020676
2  1.5217997
3 -3.0992251
```

Los resultados dicen, en el apartado `class`, que el primer y tercer caso son clasificados al grupo 1 (alumnos que logran titularse) mientras que el segundo caso es asignado al grupo 2 (no terminan la universidad). El apartado `posterior`, indica la probabilidad de que el caso pertenezca a un determinado grupo, es decir, para el primer caso la probabilidad de que se parte del grupo que termina y se titula es del 77.56% y de un 22.43% de pertenecer al grupo que no termina. Por su parte el segundo caso la probabilidad es mayor para el grupo 2 de casi el 100%. El último apartado (`x`), indica el valor de cada caso en la función discriminante. Recordando lo descrito al final del apartado E.1.3., los valores de los casos pertenecientes al grupo 1 son menores a 0, mientras que los pertenecientes al grupo 2 son mayores a 0.

E.1.5. Selección de variables

La idea del Análisis discriminante como se ha descrito es construir funciones lineales de las variables predictoras originales que discriminen entre los distintos grupos. Sin embargo, no todas las variables discriminan de la misma. Por ello, a la hora de construir las funciones lineales, no es necesario incluir a todas.

El método stepwise forward (selección hacia adelante) es un método que funciona de la siguiente manera:

i) Se incluye en el análisis la variable que tenga el mayor valor aceptable para el criterio de selección o de entrada. Como criterio general para seleccionar una variable se emplea el valor de la λ de Wilks o, de modo equivalente, del valor de su F asociada.

ii) Se evalúa el criterio de selección para las variables no seleccionadas. La variable que presenta el valor más alto para el criterio se selecciona (siempre que esté dentro de un límite).

iii) Se examinan las variables seleccionadas según un criterio de salida y se examinan también las variables no seleccionadas, para ver si cumplen el criterio de entrada. Se excluyen o se incluyen variables según cumplan los criterios de entrada y de salida.

iv) Se repite el paso (iii) hasta que ninguna variable más pueda ser seleccionada o eliminada.

```
# Procedimeinto Stepwise forward para selección de variables
# predictoras en el modelo

sc_obj <- greedy.wilks(formula = Resultado ~ Conocimiento + Aptitudes,
            data = IngresoUniversidad[1:59,-1])
sc_obj


Formula containing included variables:

Resultado ~ Aptitudes
<environment: 0x0000000003143f58>


Values calculated in each step of the selection procedure:

       vars Wilks.lambda  F.statistics.overall  p.value.overall
1 Aptitudes    0.1580097               303.7374     1.661717e-24

       F.statistics.diff  p.value.diff
                303.7374  1.661717e-24
```

Los resultados indican que utilizando únicamente la variable Aptitudes se logra la separación de ambas agrupaciones (`Resultado ~ Aptitudes`). Posteriormente muestra valores de los criterios de selección (λ de Wilks, el valor de F asociado y el p–valor) para la variable retenida. El valor de la λ de Wilks hace una comparación entre la media dentro de cada grupo con la media total sin distinguir grupos. Un valor pequeño indica que la variabilidad total de la variable se debe a la diferencia entre grupos y no a una diferencia dentro del

grupo, es decir, la variable discrimina mucho. El mismo análisis se puede hacer utilizando el estadístico F y su correspondiente p–valor donde valores menores a un determinado nivel de significancia indican el rechazo estadístico de la hipótesis nula de igualdad de varianza entre grupos. Esto significa que la variable permite hacer una adecuada discriminación entre los grupos. El resultado obtenido puede verificarse al comparar los graficos boxplot del apartado E.1.1. para ambas variables predictoras. En el grafico correspondiente a la variable *Aptitudes* no se aprecia un traslape entre los grupos contrario al grafico de la variable *Conocimiento*. Con lo cual si solo se conociera el valor de la prueba de Aptitudes se podrá saber si el candidato terminara la universidad titulado.

E.2. Discriminación para tres o más poblaciones

Tomando nuevamente como caso de aplicación el ingreso a la unidad, se desea ahora crear la "regla" que permita separar y clasificar a los estudiantes de nuevo ingreso en los tres grupos ($\mathbf{X}_1$) contenidos en la base de datos: los que logran titularse (π_1), aquellos que no terminan sus estudios (π_2), y aquellos que están en el límite de la titulación (π_3). Las variables predictoras que se utilizarán en el análisis siguen siendo las puntuaciones obtenidas en dos pruebas de ingreso.

E.2.1. Verificación de los supuestos

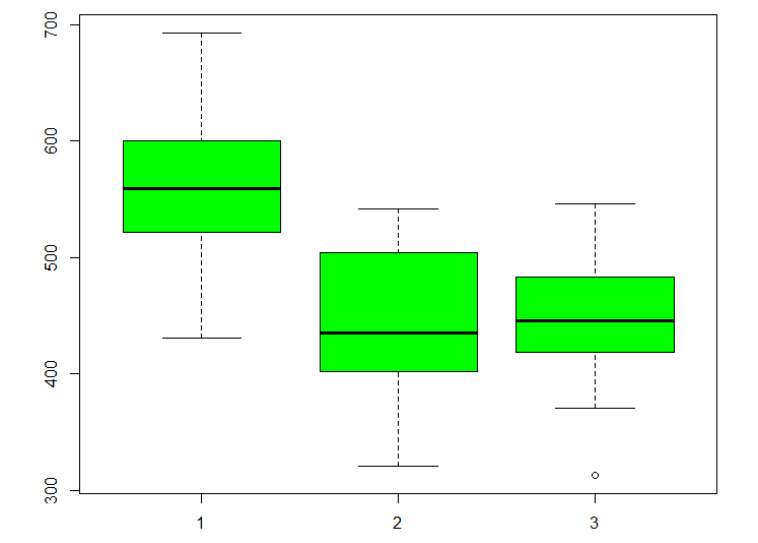
Nuevamente se realiza un análisis descriptivo ahora de los tres grupos además de verificar que se siga cumpliendo con los supuestos necesarios para decidir si es pertinente realizar el Análisis Descriptivo.

```
table(IngresoUniversidad$Resultado)

1  2  3
31 28 26

summary(IngresoUniversidad[IngresoUniversidad$Resultado == 1,
-c(1:2)])

  Conocimiento     Aptitudes
 Min.   :431.0   Min.   :2.960
 1st Qu.:522.0   1st Qu.:3.270
 Median :559.0   Median :3.390
 Mean   :561.2   Mean   :3.398
 3rd Qu.:600.5   3rd Qu.:3.540
 Max.   :693.0   Max.   :3.800


summary(IngresoUniversidad[IngresoUniversidad$Resultado == 2,
-c(1:2)])
  Conocimiento     Aptitudes
 Min.   :321.0   Min.   :2.130
 1st Qu.:404.2   1st Qu.:2.360
 Median :435.5   Median :2.470
 Mean   :447.1   Mean   :2.482
 3rd Qu.:504.2   3rd Qu.:2.578
 Max.   :542.0   Max.   :2.900
```

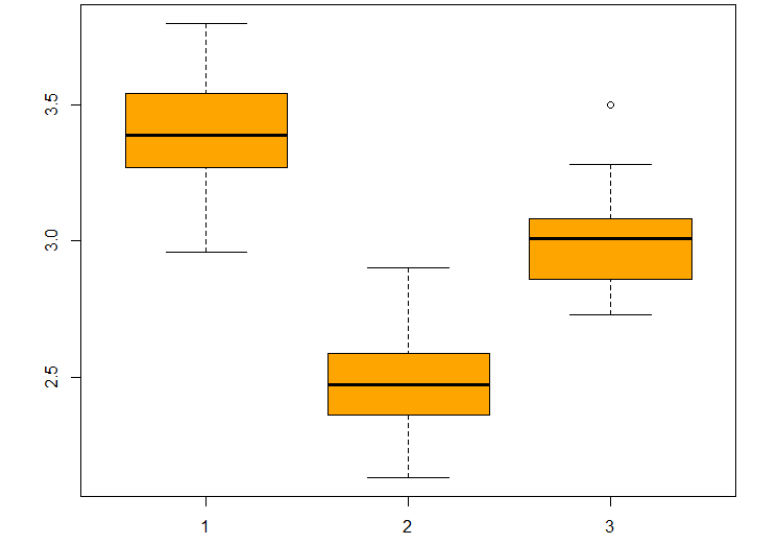
```
summary(IngresoUniversidad[IngresoUniversidad$Resultado == 3,
-c(1:2)])
```

```
  Conocimiento     Aptitudes
 Min.   :313.0   Min.   :2.730
 1st Qu.:419.0   1st Qu.:2.868
 Median :446.0   Median :3.010
 Mean   :446.2   Mean   :2.993
 3rd Qu.:480.0   3rd Qu.:3.072
 Max.   :546.0   Max.   :3.500
```

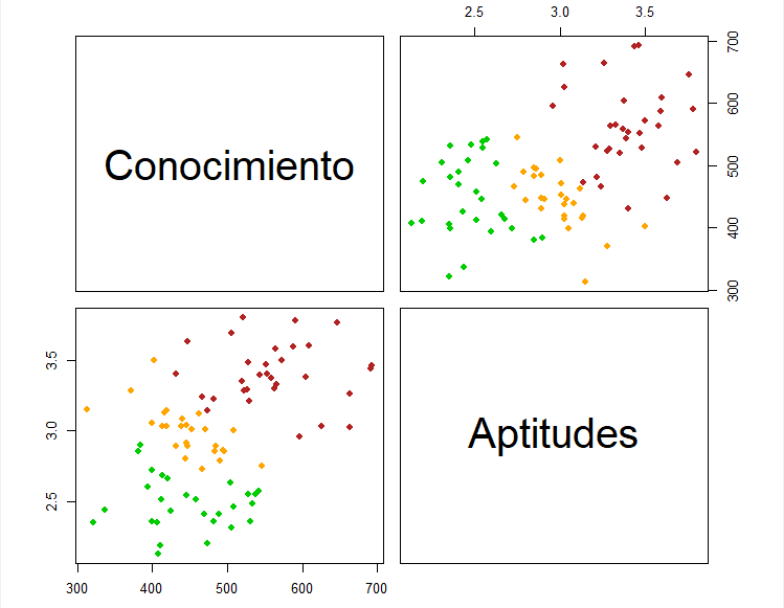
```
boxplot(IngresoUniversidad$Conocimiento[IngresoUniversidad$Resultado
== 1],IngresoUniversidad$Conocimiento[IngresoUniversidad$Resultado ==
2], IngresoUniversidad$Conocimiento[IngresoUniversidad$Resultado ==
3],col = "green")
```




```
boxplot(IngresoUniversidad$Aptitudes[IngresoUniversidad$Resultado ==
1],IngresoUniversidad$Aptitudes[IngresoUniversidad$Resultado == 2],
IngresoUniversidad$Aptitudes[IngresoUniversidad$Resultado == 3],
col = "orange")
```

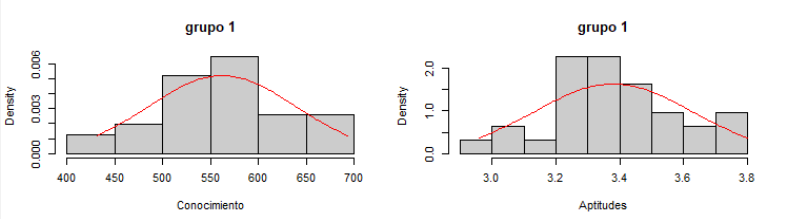


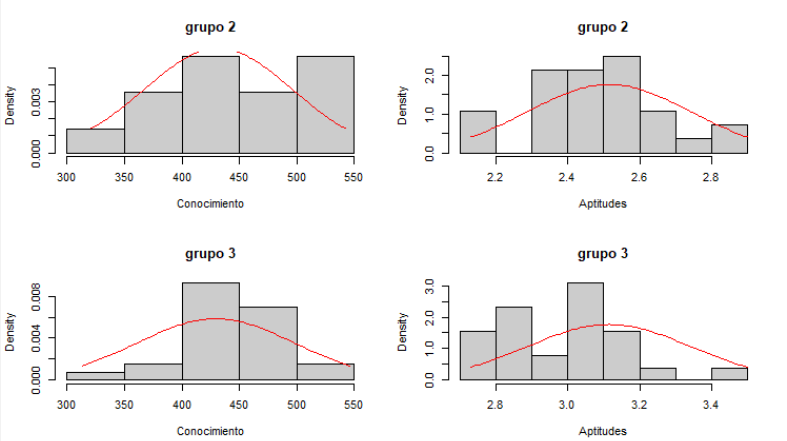
```
pairs(x = IngresoUniversidad[, c("Conocimiento","Aptitudes")],
      col = c("firebrick", "green3", "orange")[IngresoUniversidad$
      Resultado], cex.labels = 3, pch = 19)
```



A diferencia del caso con dos grupos, al introducir el tercer grupo no es tan evidente que una única variable permita hacer la separación. Se procede a verificar el supuesto de normalidad univariada.

```
# Representación mediante Histograma de cada variable para cada
# grupo
par(mfcol = c(3, 2))
for (k in 3:4) {
  j0 <- names(IngresoUniversidad)[k]
  for (i in 1:3) {
    x <- IngresoUniversidad[IngresoUniversidad$Resultado == i, j0]
    hist(x, freq = F, col = grey(0.8), main = paste("grupo", i),xlab
         = j0)
    curve(dnorm(as.matrix(x),mean(as.matrix(x)),sd(as.matrix(x))),
          min(as.matrix(x)), max(as.matrix(x)),add = T,col="red")
  }
}
par(mfcol = c(1, 1))
```






```
# Contraste de normalidad Shapiro-Wilk para cada variable en cada
# grupo
for (k in 3:4) {
  j0 <- names(IngresoUniversidad)[k]
  for (i in 1:3) {
    x <- IngresoUniversidad[IngresoUniversidad$Resultado == i, j0]
    x.test <- shapiro.test(as.matrix(x))
    x.test$data.name <- paste("grupo", i,",",j0)
    print(x.test)
  }
}
```

```
    Shapiro-Wilk normality test

data:  grupo 1 , Conocimiento
W = 0.9775, p-value = 0.7403


    Shapiro-Wilk normality test

data:  grupo 2 , Conocimiento
W = 0.94627, p-value = 0.1595


    Shapiro-Wilk normality test

data:  grupo 3 , Conocimiento
W = 0.96849, p-value = 0.5847


    Shapiro-Wilk normality test

data:  grupo 1 , Aptitudes
W = 0.97932, p-value = 0.7935


    Shapiro-Wilk normality test

data:  grupo 2 , Aptitudes
W = 0.97996, p-value = 0.8496

    Shapiro-Wilk normality test

data:  grupo 3 , Aptitudes
W = 0.93695, p-value = 0.1136
```

Se comprueba que cada una de las variables en cada uno de los grupos cumple con el supuesto al tener p–valores mayores a un nivel de significancia de 0.05.

```
library("biotools")
boxM(IngresoUniversidad[,-c(1:2)],IngresoUniversidad$Resultado)


    Box's M-test for Homogeneity of Covariance Matrices

data:  IngresoUniversidad[, -c(1:2)]
Chi-Sq (approx.) = 16.155, df = 6, p-value = 0.01295
```

Al verificar la igualdad de la covarianza entre los grupos usando la M de Box, el resultado del p–valor conduce a cuestionar el hecho que las matrices de covarianza poblacional sean iguales. La prueba muestra una significancia estadística si se considera un nivel crítico de 0.05 por lo que se consideran los grupos como diferentes y se violenta el supuesto. En este caso particular y con el propósito de mostrar el procedimiento de este análisis, siendo que el p–valor es mayor a 0.01, se procede a validar el supuesto.

```
# Matriz de correlación por grupo
round(cor(IngresoUniversidad[IngresoUniversidad$Resultado == 1,
      -c(1:2)]),2)

             Conocimiento Aptitudes
Conocimiento         1.00     -0.04
Aptitudes           -0.04      1.00


round(cor(IngresoUniversidad[IngresoUniversidad$Resultado == 2,
      -c(1:2)]),2)

             Conocimiento Aptitudes
Conocimiento          1.0      -0.1
Aptitudes            -0.1       1.0


round(cor(IngresoUniversidad[IngresoUniversidad$Resultado == 3,
-c(1:2)]),2)

             Conocimiento Aptitudes
Conocimiento         1.00     -0.66
Aptitudes           -0.66      1.00
```

Para los primeros dos grupos no se aprecia problemas de multicolinealidad aunque en el tercer grupo el valor estimado está cerca del límite. Por lo tanto se puede concluir que es factible realizar el Análisis Discriminante para este caso.

E.2.2. Estimación del modelo discriminante

```
library("MASS")

# Se crea la relación entre la variable categorica y las variables
# predictoras
IngresoUniversidad.lda <- lda(formula = Resultado ~ Conocimiento +
Aptitudes, data = IngresoUniversidad[,-1])

IngresoUniversidad.lda

Call:
lda(Resultado ~ Conocimiento + Aptitudes, data = IngresoUniversidad[,
    -1])
```

```
Prior probabilities of groups:
        1         2         3
0.3647059 0.3294118 0.3058824

Group means:
  Conocimiento Aptitudes
1     561.2258  3.398387
2     447.0714  2.482500
3     446.2308  2.992692

Coefficients of linear discriminants:
                      LD1         LD2
Conocimiento -0.008881605 -0.01431778
Aptitudes    -4.931238914  1.87356253

Proportion of trace:
   LD1    LD2
0.9668 0.0332
```

Para este caso se obtienen dos funciones discriminantes lineales que permitirán la separación de los tres grupos. Las ecuaciones quedarían expresadas de la siguiente manera:

$$\text{LD1} = -0.009 * \text{Conocimiento} - 4.931*\text{Aptitudes}$$
$$\text{LD2} = -0.014 * \text{Conocimiento} + 1.874*\text{Aptitudes}$$

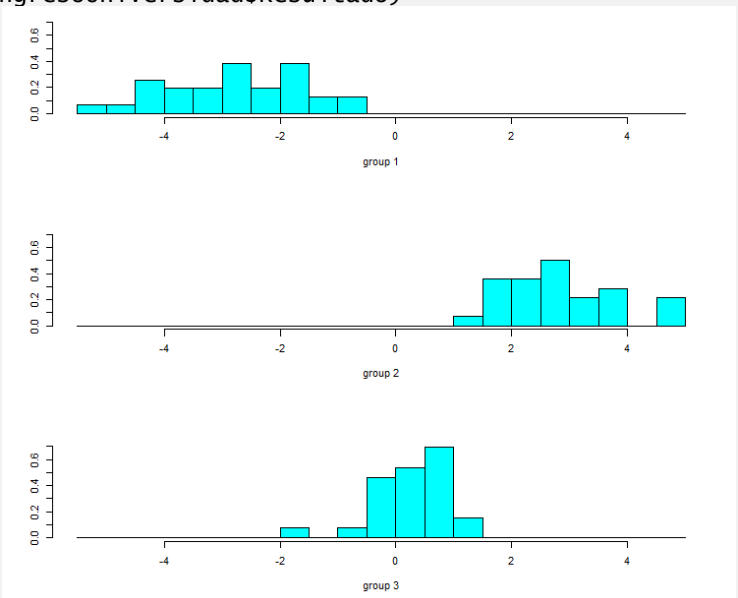
La sección denominada *proportion of trace* indica la porción de la variabilidad entre grupos que es explicado por la función discriminante. La primera función discriminante (LD1) explica el 96.68% mientras que la segunda función discriminante apenas el 3.32%.

E.1.3. Valorar la exactitud de la población

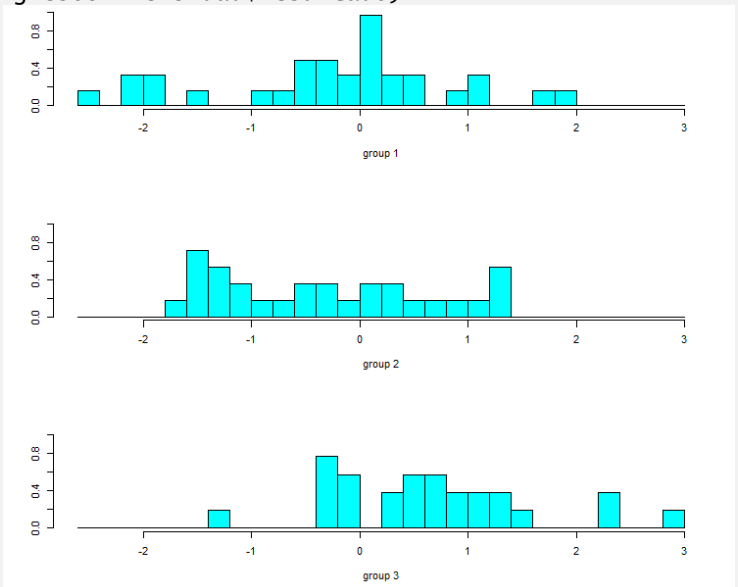
```
# Predictor de valores utilizando el modelo discriminante
IngresoUniversidad.lda.values <- predict(IngresoUniversidad.lda)


# No. de casos en cada grupo después de aplicar la función
# discriminante
table(IngresoUniversidad.lda.values$class)

1  2  3
29 27 29

# Tabla de clasificación de datos reales contra los predichos
table(IngresoUniversidad$Resultado,IngresoUniversidad.lda.
      values$class, dnn = c("Clase real", "Clase predicha"))

          Clase predicha
Clase real  1  2  3
         1 28  0  3
         2  0 26  2
         3  1  1 24


# Calculo de la tasa de error de clasificación
tasa_error <- mean(IngresoUniversidad$Resultado !=
     IngresoUniversidad.lda.values$class) * 100
paste("tasa_error (mala clasificación)=", round(tasa_error,2), "%")

[1] "tasa_error (mala clasificación)= 8.24 %"
```

Para este modelo discriminante, se tiene una tasa de error de 8.24%. Mas a detalle se contabiliza siete casos mal clasificados, de ellos tres casos que en la realidad pertenecen al grupo 1, el modelo los clasifico en el grupo 3. Dos casos que en realidad son del grupo 2, el modelo los asignó al grupo 3 mientras que 2 casos pertenecientes al grupo 3 fueron mal clasificados.

```
# Histograma de los valores predichos en cada una de las funciones
# discriminantes

# LD1
ldahist(IngresoUniversidad.lda.values $x[,1], g =
        IngresoUniversidad$Resultado)
```



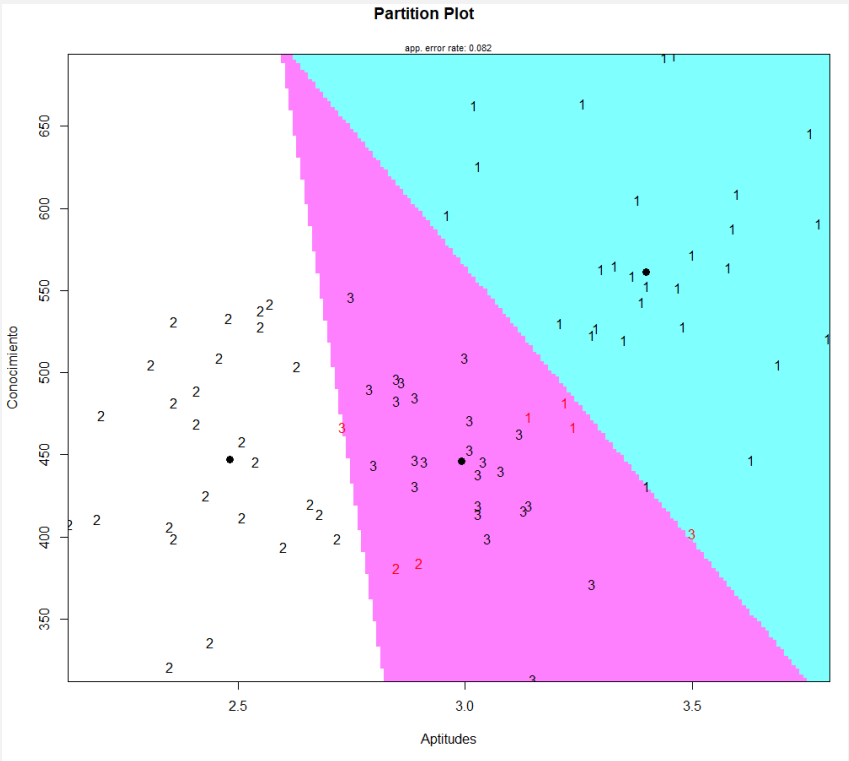

```
# LD2
ldahist(IngresoUniversidad.lda.values $x[,2], g =
        IngresoUniversidad$Resultado)
```



A fin de verificar los histogramas anteriores, se resalta que la primera función discriminante (LD1) permita una mejor discriminación de los grupos comparado a la segunda función discriminante (LD2). Esto se concluye al ver los rangos de valores que toman las observaciones en cada una de las funciones. El primer histograma, indica que el grupo 1 va de -5.5 a -0.5, el grupo 2 tiene un rango de 1 a 5 y para el grupo 3 el rango va de -2 a 1.5. Por su parte el segundo histograma despliega que el grupo 1 tiene un rango de -2.6 a 2, el grupo 2 va de -1.8 a 1.4, mientras que el grupo 3 va de -1.4 a 3. Se observa traslapes menos graves entre los rangos de los tres grupos en la LD1 que en la LD2. Esta situación confirma que la mayor diferencia entre grupos es representada por la LD1.

Una forma gráfica para representar los grupos y las funciones discriminantes se genera al utilizar el comando `partimat()`. En el grafico se representa en la dispersión de las observaciones por pares de variables indicando el grupo al que fue clasificado. Cada uno de las regiones de los grupos separadas por la función discriminante es coloreado. También son representados el centroide del grupo, aquellos casos mal clasificados se resaltan con rojo y la tasa de error (mala clasificación)

```
install.packages('klaR')
library(klaR)
IngresoUniversidad$Resultado <- factor(IngresoUniversidad$Resultado)
partimat(Resultado ~ Conocimiento + Aptitudes,data=IngresoUniversidad
         [,-1], method="lda", prec=200, name=c("Conocimiento",
         "Aptitudes"))
```



E.2.4. Clasificación de nuevos casos

Se toman los mismos casos empleados en el apartado E.1.4 con el propósito de conocer a que grupo pertenecen.

```
predict(object = IngresoUniversidad.lda, newdata = IngresoUniversidad
        _NvoCasos)

$`class`
[1] 3 2 1
Levels: 1 2 3

$posterior
             1            2          3
1 0.0608741276 1.569535e-02 0.92343052
2 0.0001268638 6.381146e-01 0.36175850
3 0.9760684317 1.565286e-08 0.02393155

$x
         LD1         LD2
1 -0.2041607 -0.04971375
2  1.5795386 -0.22329949
3 -3.1627163  2.78529377
```

El resultado indica que la primera observación es casi seguro que pertenece al grupo 3 (92.34%). La segunda nueva observación tiene mayor probabilidad de pertenecer al grupo 3 también con una probabilidad de 36.18%. Y para la tercera observación el altamente probable que pertenezca al grupo 1 (97.61%).