Análisis discriminante lineal (LDA)

(Dr. Edgar Eloy Carpio Vargas-FINESI-2024)

Contenido

1. Análisis discriminante lineal.

El Análisis Discriminante Lineal o Linear Discriminant Analysis (LDA) es un método de aprendizaje supervisado de clasificación de variables.

Para realizar un análisis discriminante la variable dependiente debe ser cualitativa con dos o más categorías de respuesta (exhaustiva y mutuamente excluyente, es decir, cada uno de los casos pertenece única y exclusivamente a un grupo) y la o las variables independientes cuantitativas (métricas). Los grupos o clases son conocidos a priori.

Este método es factible aplicarlo no solo en variables dependientes nominales, pudiendo ser esta ordinal, de intervalos e incluso de razón, en este caso la solución quedaría resuelta categorizando la variable no nominal. Ejemplo: la variable número de cigarrillos que consumen por día, puede ser categorizado como: No fumadores, poco fumadores, fumadores, fumadores empedernidos.

El objetivo del análisis discriminante, es encontrar una combinación lineal de las variables independientes que mejor diferencian o discriminar a los grupos. Una vez encontrada esa combinación lineal (función discriminante) puede ser utilizada para clasificar nuevos casos o predecir nuevos individuos. Dicho de otra manera, el Análisis Discriminante trata de establecer relaciones óptimas existentes entre las características de los individuos y sus grupos de pertenencia.

En palabras de César Pérez López, “el análisis discriminante es una técnica estadística que permite asignar o clasificar nuevos individuos dentro de grupos previamente reconocidos o definidos”.

En la clasificación discriminante hay dos enfoques: el primero está basado en la obtención de funciones discriminantes de cálculo similar a regresión lineal múltiple y el segundo emplea técnicas de correlación canónica y de componentes principales y se denomina análisis discriminante canónico.

2. Relación con otras técnicas.

El análisis discriminante es conceptualmente muy similar al análisis de varianza multivariante de un factor y es lo mismo que el análisis de regresión logística, pero a diferencia de él, solo admite variables cuantitativas. Si alguna variable independiente es categórica, es preferible utilizar la regresión logística. La principal diferencia entre el análisis discriminante y el análisis de la varianza radica en que el primero es adecuado cuando la variable dependiente es categórica, y el segundo cuando la variable dependiente es métrica. Por otro lado, en el análisis discriminante las variables independientes son métricas, en el análisis de varianza son categóricas. Con respecto a la regresión, su principal diferencia radica en que en la regresión la variable dependiente es métrica y en el análisis discriminante es categórica. El análisis discriminante, no está limitado a la obtención de una sola función como el análisis de regresión.

3. Tipos de análisis discriminante.

Se pueden clasificar dependiendo al número de categorías de la variable dependiente:

a) Análisis discriminante de dos grupos o simple, la variable dependiente tiene solo dos categorías de respuesta.

b) Análisis discriminante múltiple. la variable dependiente tiene más de dos categorías de respuesta.

4. Utilidad del análisis discriminante.

Según el objeto de la investigación.

a) Explicativos. con la intención de cuantificar la contribución relativa de cada una de las variables independientes en la clasificación correcta de los individuos considerados dentro de los distintos grupos objeto de estudio, por tanto, se intenta probar el poder discriminante de cada una de las variables, en muchos casos con la finalidad de seleccionar el subconjunto que mejor discrimina los grupos.

b) Predictivos. Encasillar a un individuo que no conocemos, dentro de un grupo a partir de los valores de las variables independientes.

c) Reclasificadores. Es decir, definidos los grupos, se desea recomponer esa partición. Este puede ser el caso cuando se desea una clasificación orientada al reconocimiento o se busca una mejor interpretación de los grupos. Así, muchas veces se realiza un análisis clúster que posteriormente se intenta corroborar por medio de un análisis discriminante.

5. Asunciones (supuestos) del análisis discriminante.

Para poder aplicar el análisis discriminante, o por lo menos para que las conclusiones sean fiables, hay que tener en cuenta los siguientes supuestos fundamentales:

1) Homogeneidad de varianzas. La matriz de covarianzas intra-grupo deben ser iguales o muy parecidos en todos los grupos objeto de estudio (dos matrices se dicen que son iguales si, y solo si, todos los elementos de las mismas coinciden). Si esto no es así, los resultados no son todo fiables, especialmente los test de significación y el proceso de clasificación. El problema es especialmente importante en el caso de que el tamaño de los diferentes grupos difiera en gran medida, ya que en el caso de que se incumpla esta restricción, se tiende a clasificar casos dentro de los grupos que tienen una matriz de covarianzas mayor. Por tanto, se debería optar por otra técnica alternativa. Para comprobar si se cumple o no esta restricción se suele recurrir al test de Box.
2) Cada uno de los grupos ha de ser una muestra procedente de una población que siga una distribución normal multivariante. (normalidad de las variables). En caso de que esto no se cumpla, se pueden producir problemas en la interpretación de las funciones discriminantes, sobre todo porque los test de significación que se aplican no son válidos, por lo que sería recomendable optar por otra técnica de análisis menos sensible a la violación de esta restricción como, por ejemplo, la regresión logística. Para comprobar la hipótesis de la normalidad multivariante los test a disposición son los de carácter gráfico. No obstante, un camino sencillo consiste en examinar primero las distribuciones de cada una de las variables individualmente consideradas (Uriel, 1995), de forma tal que, si cada variable se distribuye normalmente, las variables conjuntamente se distribuirán como una normal multivariante. Si alguna de las variables no se distribuye normalmente, hay

razones para suponer que la hipótesis de normalidad multivariante no se cumple.

3) La existencia de multicolinealidad. entre las variables independientes implica que dos o más variables están altamente correlacionadas, por lo que una variable puede ser predicha o explicada por otras, es decir tiene escasa capacidad explicativa. La multicolinealidad no supone un problema si su presencia es similar en todas las posibles muestras. En caso contrario, multicolinealidad diferente según muestras, se presenta un problema como consecuencia de que los resultados dependerán de la muestra elegida para obtener la función discriminante. (Sharma, 1996).

En la práctica, el análisis discriminante es una técnica robusta y funciona bien, aunque las restricciones anteriores no se cumplan, pero, también es aconsejable usar la regresión logística como una técnica alternativa.

La varianza del predictor es igual en todas las clases de la variable respuesta. En el caso de múltiples predictores, la matriz de covarianza es igual en todas las clases. Si esto no se cumple se recurre a Análisis Discriminante Cuadrático (QDA).

Cuando la condición de normalidad no se cumple, el LDA pierde precisión, pero aun así puede llegar a clasificaciones relativamente buenas. Using discriminant analysis for multi-class classification: an experimental investigation (Tao Li, Shenghuo Zhu, Mitsunori Ogihara).

6. Recomendaciones respecto a la muestra.

a) La muestra debe ser representativa de cada uno de los grupos que estén constituidos a priori. Sin embargo, no es necesario que el tamaño de la muestra de cada grupo sea el mismo.

b) Las variables deberán ser elegidas de manera que puedan definir y discriminar los grupos, por tanto, deberán ser lo más independientes posible unas de otras.

c) El tamaño mínimo de la muestra (recomendado) es de cinco observaciones para cada variable independiente, pero, muchos estudios sugieren un ratio de 20 observaciones por cada variable independiente.

d) El grupo más pequeño en miembros debería exceder al número de variables independientes.

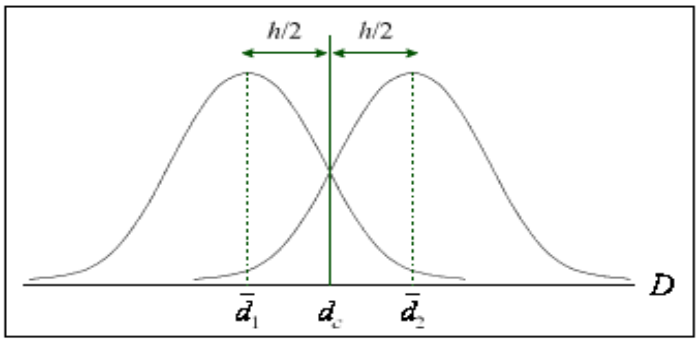
e) Un caso será excluido del análisis si no se tiene información sobre el mismo acerca de la variable que define el grupo de pertenencia o de alguna del resto de las variables utilizadas como predictoras. No obstante, conviene tomar una serie de precauciones antes de eliminar casos del análisis. En primer lugar, porque puede que nos quedemos con muy pocos, por lo que las posibilidades de generalizar los resultados obtenidos disminuirían considerablemente; en segundo lugar, porque de los casos de los que se carece información sobre alguna de las variables independientes difieren en aquellos sobre los que se tiene toda la información, los resultados estarán sesgados. Si la ausencia de información en una variable para algunos casos se debiera a alguna característica distintiva de estos (por ejemplo, clase social, nivel de estudios, raza, religión, etc.), sería aconsejable, antes de eliminar los casos, prescindir de tales variables.

7. Planteamiento o formulación del problema

La mayor utilidad de una función discriminante radica en su capacidad para clasificar nuevos casos. Ahora bien, la clasificación de casos es algo muy distinto de la estimación de la función discriminante. De hecho, una función perfectamente estimada puede no pasar de una pobre capacidad clasificatoria.

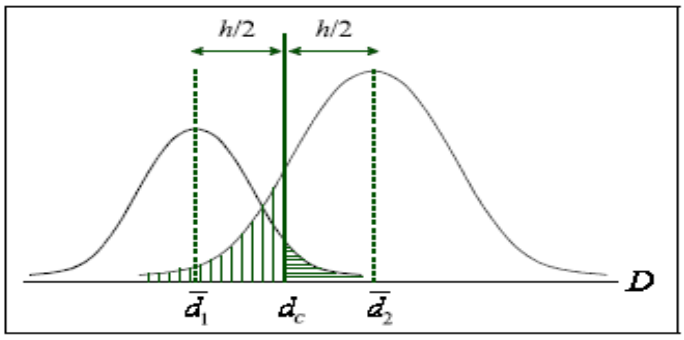
Una manera de clasificar los casos consiste en calcular la distancia existente entre los centroides de ambos grupos y situar un punto de corte dc equidistante de ambos centroides (figura 1). A partir de ese momento, los casos cuyas puntuaciones discriminantes sean mayores que el punto de corte dc serán asignados al grupo superior y los casos cuyas puntuaciones discriminantes sean menores que el punto de corte dc serán asignados al grupo inferior.

Figura 1. Utilización de un punto de corte equidistante de ambos centroides



Esta regla de clasificación tiene un serio inconveniente: sólo permite distinguir entre dos grupos y es difícilmente aplicable al caso de más de dos grupos. Además, no tiene en cuenta que los grupos pueden tener distinto tamaño. Si los tamaños muestrales son muy desiguales, la situación real será más parecida a la que muestra la figura 2. En esta figura puede verse con claridad que, si utilizamos el punto de corte dc como punto de clasificación, la proporción de casos mal clasificados en el grupo de menor tamaño (zona rayada horizontalmente) será mucho menor que en el grupo de mayor tamaño (zona rayada verticalmente).

Figura 2. Punto de corte equidistante de ambos centroides.



Por tanto, con tamaños desiguales es preferible utilizar una regla de clasificación que desplace el punto de corte hacia el centroide del grupo de menor tamaño buscando igualar los errores de clasificación. Para calcular este punto de corte podemos utilizar una distancia ponderada:

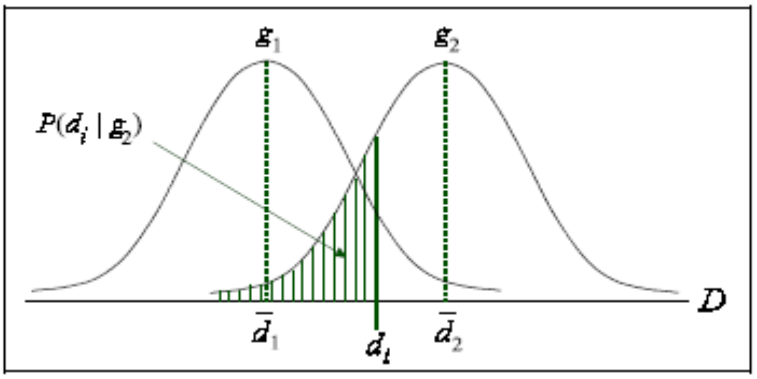
$$\bar{d}_c = \frac{n_1\bar{d}_1 + n_2\bar{d}_2}{n_1 + n_2}$$

Fukunaga y Kessell (1973) y Glick (1978) han propuesto una regla de clasificación basada en la teoría bayesiana. Esta otra regla permite incorporar fácilmente la información relativa al tamaño de los grupos y, además, es extensible al caso de más de dos grupos.

Las probabilidades a priori ofrecen alguna información sobre la representatividad de los casos, pero no ofrecen información concreta sobre un caso en particular. Además, las probabilidades a priori no tienen en cuenta que las probabilidades de aparición de las variables independientes en cada grupo pueden no ser simétricas.

Por supuesto, siempre es posible aprovechar la información adicional que proporciona saber a qué grupo pertenece cada caso. Si asumimos que las puntuaciones discriminantes se distribuyen normalmente, podemos calcular la probabilidad asociada a un caso (es decir, la probabilidad que queda por encima o por debajo de ese caso) en cada uno de los grupos utilizados en el análisis. Esto es lo que se conoce como probabilidad condicional: $P(D>d_i/G=g_k)$ o, simplemente, $P(d_i|g_k)$. La probabilidad condicional de una puntuación discriminante puede calcularse mediante tablas de probabilidad asintótica o a partir de los cuantiles observados (ver figura 3).

Figura 3. Probabilidad condicional de la puntuación discriminante di en el grupo 2



Una puntuación discriminante tiene asociadas tantas probabilidades condicionales como grupos hay en el análisis. Esas probabilidades condicionales indican cómo es de probable una puntuación concreta en cada uno de los grupos. Pero sólo son útiles cuando se conoce a qué grupo pertenece un caso. Cuando se desea clasificar un caso nuevo (del que, obviamente se desconoce a qué grupo pertenece), es necesario comparar las probabilidades condicionales que le corresponden en cada uno de los grupos del análisis. Por ello, para clasificar un caso nuevo, es más apropiado utilizar las probabilidades a posteriori, es decir, las probabilidades de pertenecer a cada uno de los grupos, dado que a ese caso le corresponde una determinada puntuación discriminante, es decir: $P(G=g_k/D=d_i)$ o, simplemente, $P(g_k/d_i)$. Estas probabilidades a posteriori se obtienen utilizando el teorema de Bayes:

$$p(g_k/d_i) = \frac{p(d_i/g_k)p(g_k)}{\sum_{k=1}^{g} p(d_i/g_k)p(g_k)}$$

La sumatoria del denominador posee tantos términos como grupos (no hay límite en el número de grupos). Con esta regla de clasificación, los casos nuevos son clasificados en el grupo al que corresponde mayor probabilidad a posteriori.

Aunque en la estimación de las probabilidades a priori es habitual utilizar los tamaños de los grupos, la aplicación del teorema de Bayes permite manipular esas probabilidades y asignarles un valor arbitrario (para reflejar mejor la composición de la población, para compensar el coste de una clasificación errónea, etc.). La manipulación de las probabilidades a priori hace que se desplace el punto de clasificación. Si se asigna igual probabilidad a priori a todos los grupos, el punto de corte para la clasificación será equidistante de todos ellos; si se aumenta la probabilidad a priori de un grupo, el punto de corte para la clasificación se alejará de su centroide.

Una forma más de determinar el punto de corte óptimo para la clasificación consiste en la curva COR (curva característica del receptor ideal).

Ninguno de los procedimientos mencionados valora el coste de la clasificación errónea de los sujetos: todos ellos asumen igual coste para los aciertos y los errores en todos los grupos. Si existen costes diferenciales para cada tipo de acierto y para cada tipo de error, será necesario establecer el punto de corte mediante otro tipo de procedimientos más característicos de la Teoría de la toma de decisiones.

LDA busca las direcciones de máxima separación de clases, ejemplo:

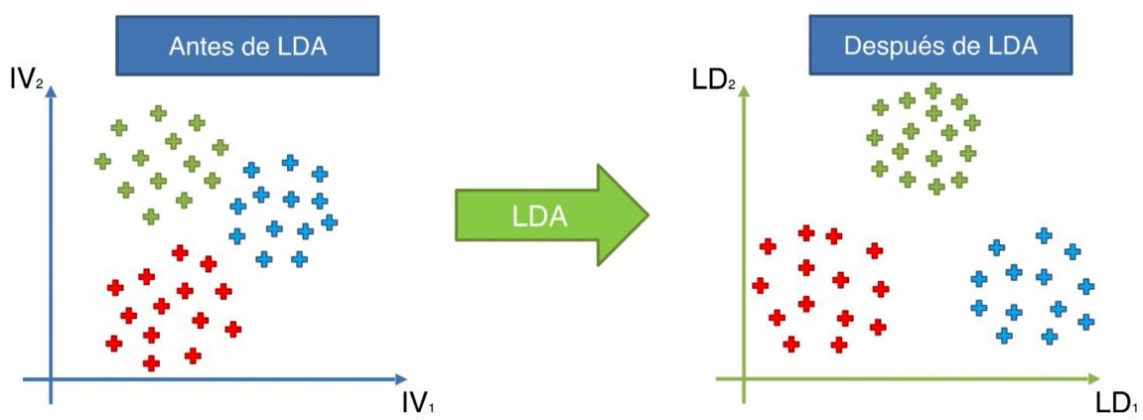


Figura 4. Ante y después de una máxima separación de clases

8. Ventajas de LDA frente a una logística múltiple.

Si bien existen extensiones de la regresión logística para múltiples clases, el LDA presenta una serie de ventajas:

- Si las clases están bien separadas, los parámetros estimados en el modelo de regresión logística son inestables. El método de LDA no sufre este problema.
- Si el número de observaciones es bajo y la distribución de los predictores (variables independentes) es aproximadamente normal en cada una de las clases, LDA es más estable que la regresión logística. La normalidad es un requisito, pero no es indispensable.

Cuando se trata de un problema de clasificación con solo dos niveles, ambos métodos suelen llegar a resultados similares.

9. El proceso de un LDA suele resumirse en 6 pasos:

- Disponer de un conjunto de datos en el que se conoce a que grupo pertenece cada observación. Puede disponer también data *training y data test.*
- Calcular las probabilidades previas (*prior probabilities*): la proporción esperada de observaciones que pertenecen a cada grupo.
- Determinar si la varianza o matriz de covarianzas es homogénea en todos los grupos. De esto dependerá que se emplee *LDA* o *QDA*.
- Estimar los parámetros necesarios para las funciones de probabilidad condicional, verificando que se cumplen las condiciones para hacerlo.
- Calcular el resultado de la función discriminante. El resultado determina a qué grupo se asigna cada observación.
- Utilizar validación cruzada (*cross-validation*) para estimar las probabilidades de clasificaciones erróneas.

10. Cálculo de las funciones discriminantes (extracción de las funciones)

La discriminación entre los *q* grupos se realiza mediante el cálculo de unas funciones matemáticas denominadas funciones discriminantes. Existen varios procedimientos para calcularlas siendo el procedimiento de Fisher uno de los más utilizados.

Dos aproximaciones a LDA: Bayes y Fisher

Existen varios enfoques posibles para realizar un LDA. La aproximación descrita anteriormente está basada en el clasificador de Bayes, y utiliza todas las variables originales para calcular las probabilidades posteriores de que una observación pertenezca a cada grupo.

Antes de que el clasificador de Bayes fuese introducido en el LDA, Fisher propuso una aproximación en la que el espacio p-dimensional (donde p es el número de predictores originales) se reduce a un subespacio de menos dimensiones formado por las combinaciones lineales de los predictores que mejor explican la separación de las clases. Una vez encontradas dichas combinaciones se realiza la clasificación en este subespacio. Fisher definió como subespacio óptimo a aquel que maximiza la distancia entre grupos en términos de varianza. Los términos de discriminante lineal de Fisher y LDA son a menudo usados para expresar la misma idea, sin embargo, el artículo original de Fisher describe un discriminante ligeramente diferente, que no hace algunas de las suposiciones del LDA como la de una distribución normal de las clases o covarianzas iguales entre clases.

La aproximación de Fisher se puede ver como un proceso con dos partes:

- Reducción de dimensionalidad: Se pasa de p variables predictoras originales a k combinaciones lineales de dichos predictores (variables discriminantes) que permiten explicar la separación de los grupos, pero con menos dimensiones ($k < p$).
- Clasificación de las observaciones empleando las variables discriminantes.

Los resultados de clasificación obtenidos mediante el método de Fisher son iguales a los obtenidos por el método de Bayes cuando:

- En el método de Bayes se asume que la matriz de covarianzas en igual en todos los grupos y se emplea como estimación la pooled within-class covariance matrix.

- En el método de Fisher, todos los discriminantes lineales se utilizan para la clasificación. El número máximo de discriminantes obtenido tras la reducción de dimensionalidad es número grupos-1.

Bayes Optimality in Linear Discriminant Analysis Onur C. Hamsici and Aleix M. Martinez Generalizing Fisher's linear discriminant analysis via the SIR approach, Chapter 14

http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema1dm.pdf

Procedimiento Discriminante de Fisher.

El procedimiento de Fisher toma como funciones discriminantes, combinaciones lineales de las variables clasificadoras de la forma:

$$D = u_1X_1 + u_2X_2 + \ \ldots\ldots + u_kX_k$$

Se trata de obtener los coeficientes de ponderación u_j. Si se considera que existe n observaciones, la función sería:

$$D_i = u_1X_{1i} + u_2X_{2i} + \ \ldots\ldots + u_kX_{ki} \qquad i = 1,2,\ldots,n$$

D_i es la puntuación discriminante correspondiente a la observación i-esima.

Expresando las variables explicativas en desviaciones respecto a la media, D_i lo estará y la relación anterior se puede expresar en forma matricial como sigue:

Sean $\{d_{gk}\ ;\ k = 1,\ldots,n_g;\ g = 1,\ldots,q\}$, los valores de la variable D en cada uno de los q grupos donde d_{gk} denota el valor de D en la k-ésima observación del g-ésimo grupo.

Sean,

$$\bar{d}_g = \frac{\sum_{k=1}^{n_g} d_{gk}}{n_g};\ \ g = 1,\ldots,q$$

Las medias muestrales de la variable D en cada uno de los q grupos y sea

$\bar{d} = \frac{\sum_{g=1}^{q}\sum_{k=1}^{n_g} d_{gk}}{n}$ la media de la variable D.

El procedimiento de Fisher determina el vector **u** que maximiza el cociente:

$$\frac{variabilidad\ entre\ grupos}{variabilidad\ intra\ grupos} = \frac{\dfrac{\sum_{g=1}^{q} n_g\left(\bar{d}_g - \bar{d}\right)^2}{q-1}}{\dfrac{\sum_{g=1}^{q}\sum_{k=1}^{n_g}\left(d_{gk} - \bar{d}_g\right)^2}{n-q}} = \frac{\mathbf{u'Bu}}{\mathbf{u'Wu}} \qquad \frac{n-q}{q-1}$$

donde:

$$W = \sum_{g=1}^{G}\sum_{k=1}^{n_j}\left(y_{gk} - \bar{y}_g\right)\left(y_{gk} - \bar{y}_g\right) = \sum_{g=1}^{G} W_g$$

$$\begin{bmatrix} \sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_{1gk}-\bar{y}_{1g}\right)^2 & \cdots & \sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_{1gk}-\bar{y}_{1g}\right)\left(y_{Kgk}-\bar{y}_{Kg}\right) \\ \cdots & \cdots & \cdots \\ \sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_{Kgk}-\bar{y}_{Kg}\right)\left(y_{1gk}-\bar{y}_{1g}\right) & \cdots & \sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_{Kgk}-\bar{y}_{Kg}\right)^2 \end{bmatrix}$$

es la matriz de suma de cuadrados intra-grupos

$B = \sum_{g=1}^{G} n_g(y_g - \bar{y})(y_g - \bar{y})$=

$$\begin{bmatrix} \sum_{g=1}^{G} n_g\left(\bar{y}_{1g}-\bar{y}_1\right)^2 & \cdots & \sum_{g=1}^{G} n_g\left(y_{1g}-\bar{y}_1\right)\left(y_{Kg}-\bar{y}_K\right) \\ \cdots & \cdots & \cdots \\ \sum_{g=1}^{G} n_g\left(y_{Kg}-\bar{y}_K\right)\left(y_{1g}-\bar{y}_1\right) & \cdots & \sum_{g=1}^{G} n_g\left(\bar{y}_{Kg}-\bar{y}_K\right)^2 \end{bmatrix}$$

Es la matriz de suma de cuadrados inter-grupos.

Se impone, además, la condición de normalización **u'Wu = 1**

La solución viene dada por el vector propio $\mathbf{u_1}$ de $\mathbf{W^{-1}B}$ asociado al mayor valor propio λ_1 de esta matriz.

En general, si se quieren calcular *r* funciones discriminantes con varianza 1, y que sean incorreladas entre sí, es decir, que verifiquen que $\mathbf{u_i'Wu_j} = \delta_{ij}$; i,j = 1,…,*r*, se obtienen como soluciones los *r* vectores propios de $\mathbf{W^{-1}B}$ asociados a los *r* mayores valores propios de esta matriz $\lambda_1 \geq \ldots \geq \lambda_1 > 0$. A las funciones $D_i = \mathbf{u_i'Y}$; *i=1,…,r* se les llama funciones discriminantes canónicas o funciones discriminantes de Fisher.

Observación

Si *r* es el número de funciones discriminantes se tiene que $\mathbf{WD} = \mathbf{I_r}$ y $\mathbf{BD} = diag(\lambda_1,\ldots,\lambda_r)$ donde **WD** y **BD** son las matrices **W** y **B** calculadas utilizando las puntuaciones discriminantes. Se sigue que:

$$\lambda_i = \sum_{g=1}^{q} n_g\left(\bar{d}_g^i - \bar{d}^i\right)^2; \quad i = 1, \ldots r$$

donde $\{\bar{d}_g^i;\ g=1,..,q\}$ son las puntuaciones medias de la i-ésima función discriminante en los *q* grupos y $\bar{d}^i$ es la puntuación media total.

Por lo tanto, los valores propios $\{\lambda_i$; i=1,...,r$\}$ miden el poder de discriminación de la i-ésima función discriminante de forma que si $\lambda_i = 0$ la función discriminante no tiene ningún poder discriminante. Dado que el rango de la matriz $\mathbf{W^{-1}B}$ es a lo más min{*q-1,p*} el número máximo de funciones discriminantes que se podrán calcular será igual a min{*q-1,p*}.

Lambda de Wilks para las funciones discriminantes. (Evaluación de la significación de las funciones discriminantes)

Con la función ya desarrollada, el siguiente paso sería determinar si esta es verdaderamente significativa, es decir, si es capaz de discriminar entre grupos. Así pues la hipótesis nula de que la media de todas las funciones discriminantes en todos los grupos es igual debe ser probada estadísticamente.

H_o: la media de todas las funciones discriminantes en todos los grupos es igual

$$H_0: \frac{Y}{G_i} \sim N_p(\mu_i, \sum\ ); i = 1, .. q\ con\ \mu_1 = \cdots = \mu_q,$$

$$\Leftrightarrow H_0: \lambda_1 = \cdots = \lambda_{min\{q-1,p\}} = 0$$

H_a: la media de todas las funciones discriminantes en todos los grupos no son iguales.

Usando los valores de la "chi cuadrado". Si p es menor que el nivel de significancia se rechaza la hipótesis nula.

El estadístico viene dada por,

$$\Lambda = \frac{|W|}{|W+B|} = \frac{1}{\prod_{i=1}^{\min\,(q-1,p)}(1+\lambda_i)}$$

Que indica la proporción total de la varianza en los resultados discriminantes no explicada por las diferencias entre grupos. ***Su valor oscila entre 0 y 1, cuanto más cerca de 0 esté, mayor es el poder discriminante de las variables consideradas y cuanto más cerca de 1, menor es dicho poder y no será adecuado para construir funciones discriminantes***.

Este estadístico tiene una distribución lambda de Wilks con *p*, *q-1* y *n-q* grados de libertad.

M de box.

Prueba M de Box para el contraste de la hipótesis nula de igualdad de las matrices de varianzas-covarianzas poblacionales. Uno de los supuestos del análisis discriminante es que todos los grupos proceden de la misma población y, más concretamente, que las matrices de varianzas-covarianzas poblacionales correspondientes a cada grupo son iguales entre sí. El estadístico M de box toma la forma:

$$M = (n-g)log|S| - \sum_{j=1}^{g}(n_j - 1)log\left|S^{(j)}\right|$$

Donde S es la matriz de varianzas-covarianzas combinada, $S^{(j)}$ es la matriz de varianzas-covarianzas del j-ésimo grupo, *n* es el número total de casos, n_j es el número de casos en el j-ésimo grupo y *g* es el número de grupos. El estadístico M carece de distribución muestral conocida, pero puede transformase en un estadístico F e interpretarse como tal (muchos analistas critican el uso de este estadístico por ser demasiado sensible a pequeñas desviaciones de la normalidad multivariante y a tamaños muestrales grandes, tendiendo a ser conservador).

Correlación canónica y autovalores asociados a una función discriminante.

La i-ésima correlación canónica viene dada por:

$$CRi = \sqrt{\frac{\lambda_i}{1+\lambda_i}} \quad i = 1, \dots r$$

y mide, en términos relativos, el poder discriminante de la i-esima función discriminante ya que es el porcentaje de la variación total en dicha función que es explicada por las diferencias entre los grupos.

Toma valores entre 0 y 1 de forma que, cuanto más cerca de 1 esté su valor, mayor es la potencia discriminante de la i-esima función discriminante.

Mediante el estadístico lambda hemos comprobado que la información que aportará cada uno de las tres funciones es estadísticamente significativa, pero no conocemos que parte de la información será atribuible a cada una de ellas. La correlación canoníca y el autovalor asociado a una función son dos medidas relacionadas con Lambda de Wilks, que permiten evaluar la información que aportara cada función discriminante en particular. La *correlación canoníca* mide las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones totales sin distinguir grupos. El *autovalor* mide las desviaciones de las puntuaciones discriminantes entre los grupos respecto a las desviaciones dentro de los grupos. En ambos casos, si el valor obtenido es grande (en el caso particular de la correlación canoníca, si es próximo a 1) la dispersión será debida a las diferencias entre grupos y, en consecuencia, la función discriminará mucho los grupos.

Teorema de Bayes para clasificación

Considérense dos eventos A y B, el teorema de Bayes establece que la probabilidad de que B ocurra habiendo ocurrido A ($B|A$) es igual a la probabilidad de que A y B ocurran al mismo tiempo (AB) dividida entre la probabilidad de que ocurra A.

$$P(B|A) = \frac{P(AB)}{P(A)}$$

Supóngase que se desea clasificar una nueva observación en una de las K clases de una variable cualitativa Y, siendo $K \geq 2$, a partir de un solo predictor X. Se dispone de las siguientes definiciones:

- Se define como overall, prior probability o probabilidad previa (π_k) la probabilidad de que una observación aleatoria pertenezca a la clase k.
- Se define (X) ≡ (X=x|Y=k) como la función de densidad de probabilidad condicional de X para una observación que pertenece a la clase k. Cuanto mayor sea (X) mayor la probabilidad de que una observación de la clase k adquiera un valor de $X \approx x$.
- Se define como posterior probability o probabilidad posterior (Y=k|X=x) la probabilidad de que una observación pertenezca a la clase k siendo x el valor del predictor.

Aplicando del teorema de Bayes se pueden conocer la posterior probability para cada clase:

$$P\ (pertenecer\ a\ la\ clase\ k|\ valor\ x\ observado) = \frac{p(pertenecer\ a\ la\ clase\ k\ y\ observar\ x)}{P(observar\ x)}$$

Si se introducen los términos, definidos anteriormente, dentro la ecuación se obtiene:

$$P(Y = k/X = x) = \frac{\pi_k P(X = x/Y = k)}{\sum_{j=1}^{k} \pi_j P(X = x/Y = j)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^{k} \pi_j f_j(x)}$$

La clasificación con menor error (clasificación de Bayes) se consigue asignando la observación a aquel grupo que maximice la posterior probability. Dado que el denominador $\sum_{j=1}^{k} \pi_j f_j(x)$ es igual para todas las clases, la norma de clasificación es equivalente a decir que se asignará cada observación a aquel grupo para el que (x) sea mayor.

Para que la clasificación basada en Bayes sea posible, se necesita conocer la probabilidad poblacional de que una observación cualquiera pertenezca a cada clase (π_k) y la probabilidad poblacional de que una observación que pertenece a la clase k adquiera el valor x en el predictor ($f_k(X) \equiv P(X = x|Y = k)$). En la práctica, raramente se dispone de esta información, por lo que los parámetros tienen que ser estimados a partir de la muestra. Como consecuencia, el clasificador LDA obtenido se aproxima al clasificador de Bayes, pero no es igual.

Estimación de π_k y $f_k(X)$

La capacidad del LDA para clasificar correctamente las observaciones depende de cómo de buenas sean las estimaciones de π_k y (X). Cuanto más cercanas al valor real, más se aproximará el clasificador LDA al clasificador de Bayes. En el caso de la prior probability (π_k) la estimación suele ser sencilla, la probabilidad de que una observación cualquiera pertenezca a la clase k es igual al número de observaciones de esa clase entre el número total de observaciones $\hat{\pi}_k = \frac{n_k}{N}$

La estimación de (X) no es tan directa y para conseguirla se requiere de ciertas asunciones. Si se considera que (X) se distribuye de forma normal en las K clases, entonces se puede estimar su valor a partir de la ecuación:

$$f_k(X) = P(Y = k/X = x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

Donde μ_k y $\sigma^2{}_k$ son la media y la varianza para la clase k.

Si además se asume que la varianza es constante en todos los grupos $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$, entonces, la sumatoria $\sum_{j=1}^{k} \pi_j f_j(x)$ se simplifica en gran medida permitiendo calcular la posterior probability según la ecuación:

$$P(Y = k/X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x - \mu_j\right)^2\right)}$$

Esta ecuación se simplifica aún más mediante una transformación logarítmica de sus dos términos:

$$\hat{\delta}_k(x) = log(P(Y = k|X = x)) = x\,\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

El término lineal en el nombre Análisis discriminante lineal se debe al hecho de que la función discriminatoria es lineal respecto de X.

En la práctica, a pesar de tener una certeza considerable de que X se distribuye de forma normal dentro de cada clase, los valores $\mu_1 \ldots \mu_k$, $\pi_1 \ldots \pi_k$ y σ^2 se desconocen,

por lo que tienen que ser estimados a partir de las observaciones. Las estimaciones empleadas en LDA son:

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_1} x_i \qquad \hat{\sigma}_k = \frac{1}{N-K}\sum_{k=1}^{k}\sum_{i:y_1}(x_i - \hat{\mu}_k)^2 \qquad \hat{\pi}_k = \frac{n_k}{N}$$

$\hat{\mu}_k$ es la media de las observaciones del grupo k, $\hat{\sigma}_k$ es la media ponderada de las varianzas muestrales de las K clases y $\hat{\pi}_k$ la proporción de observaciones de la clase k respecto al tamaño total de la muestra.

La clasificación de Bayes consiste en asignar cada observación X = x a aquella clase para la que $(Y = k|X = x)$ sea mayor. En el caso particular de una variable cualitativa Y con solo dos niveles, se puede expresar la regla de clasificación como un ratio entre las dos posterior probabilities. Se asignará la observación a la clase 1 si $\frac{(Y=1/X=x)}{(Y=2/X=x)} > 1$, y a la clase 2 si es menor.

En este caso particular el límite de decisión de Bayes viene dado por $x = \frac{\mu_1+\mu_2}{2}$

La siguiente imagen muestra dos grupos distribuidos de forma normal con medias $\mu_1 = -1.25$ y $\mu_2 = 1.25$ y varianzas $\sigma_1^2 = \sigma_2^2 = 1$. Dado que se conoce el valor real de las medias y varianzas poblacionales (esto en la realidad no suele ocurrir), se puede calcular el límite de decisión de Bayes $x = \frac{-1.25+1.25}{2} = 0$ (línea discontinua). La línea discrimina los valores de la variable dependiente.

```
library(ggplot2)
ggplot(data.frame(x = c(-4, 4)), aes(x)) +
  stat_function(fun = dnorm, args = list(mean = -1.25, sd = 1), color="firebrick") +
  stat_function(fun = dnorm, args = list(mean = 1.25, sd = 1), color = "green3") +
  geom_vline(xintercept = 0, linetype = "longdash") + theme_bw()
```

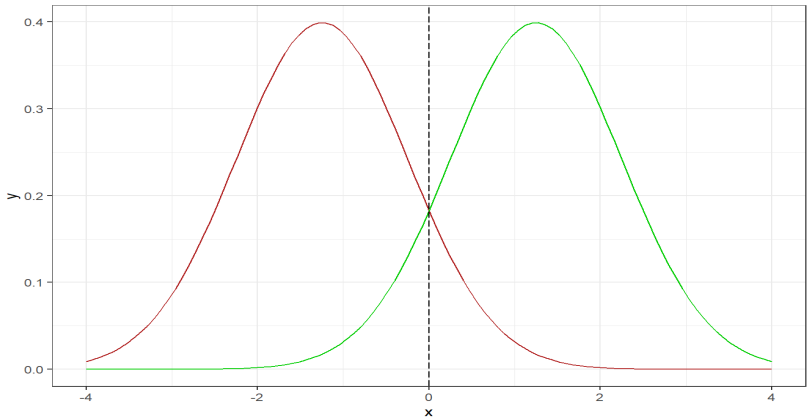


Figura. límite de decisión de Bayes $x = 0$ (línea discontinua)

Si en lugar de conocer la verdadera distribución poblacional de cada grupo solo se dispone de muestras, escenario que suele ocurrir en los casos reales, el límite de decisión LDA se aproxima al verdadero límite de decisión de Bayes, pero no es exacto. Cuanto más representativas sean las muestras mejor la aproximación.

```
set.seed(6911)
library(ggplot2)
grupo_a <- rnorm(n = 30, mean = -1.25, sd = 1)
grupo_b <- rnorm(n = 30, mean = 1.25, sd = 1)
datos <- data.frame(valor = c(grupo_a, grupo_b), grupo = rep(c("A", "B"),
                                                             each = 30))
ggplot(data = datos, aes(x = valor, fill = grupo)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  geom_vline(xintercept = 0, linetype = "longdash") +
  geom_vline(xintercept = (mean(grupo_a) + mean(grupo_b))/2) +
  geom_text(aes(+1.2, 9, label = "Limite decision LDA")) +
  geom_text(aes(-1.2, 10, label = "Limite decision Bayes")) +
  theme_bw()
```

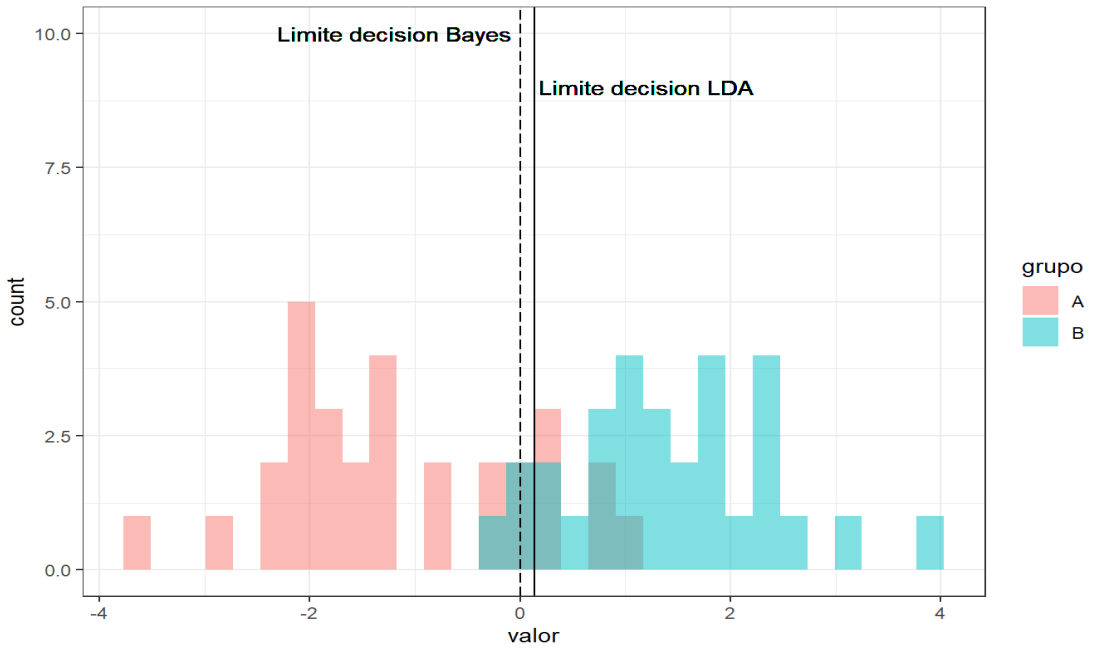


Figura … Límite de decisión LDA se aproxima al verdadero límite de decisión de Bayes, pero no es exacto.

Extensión del LDA para múltiples predictores

Los conceptos anteriormente descritos empleando un único predictor pueden generalizarse para introducir múltiples predictores en el modelo. La diferencia reside en que X, en lugar de ser un único valor, es un vector formado por el valor de p predictores $X = (X_1, X_2, \ldots, X_3)$ y que, en lugar de proceder de una distribución normal, procede de una distribución normal multivariante.

Un vector sigue una distribución k-normal multivariante si cada uno de los elementos individuales que lo forman sigue una distribución normal y lo mismo para toda combinación lineal de sus k elementos. Las siguientes imágenes muestran representaciones gráficas de distribuciones normales multivariante de 2 elementos (distribución normal bivariante).

```
# Codigo obtenido de:
# http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-11/code-5/
# R code to create bivariate figure
mu1 <- 0  # set mean x1
mu2 <- 0  # set mean x2
s11 <- 10 # set variance x1
s22 <- 10 # set variance x2
s12 <- 15 # set covariance x1 and x2
rho <- 0.5 # set correlation coefficient  x1 and x2
x1 <- seq(-10,10,length = 41) # generate vector  x1
x2 <- x1 # copy x1 to x2
f <- function(x1,x2) # multivariate function
{
  term1 <- 1/(2*pi*sqrt(s11*s22*(1-rho^2)))
  term2 <- -1/(2*(1-rho^2))
  term3 <- (x1-mu1)^2/s11
  term4 <- (x2-mu2)^2/s22
  term5 <- -2*rho*((x1-mu1)*(x2-mu2))/(sqrt(s11)*sqrt(s22))
  term1*exp(term2*(term3+term4-term5))
}
z <- outer(x1,x2,f) # calculate density values
persp(x1, x2, z, # 3-D plot
      main = "Distribucion multivariante con dos predictores",
      col ="lightgreen",
      theta = 30, phi = 20,
      r = 50,
```

```
    d = 0.1,
    expand = 0.5,
    ltheta = 90, lphi = 180,
    shade = 0.75,
    ticktype = "simple",
    nticks = 5)
```

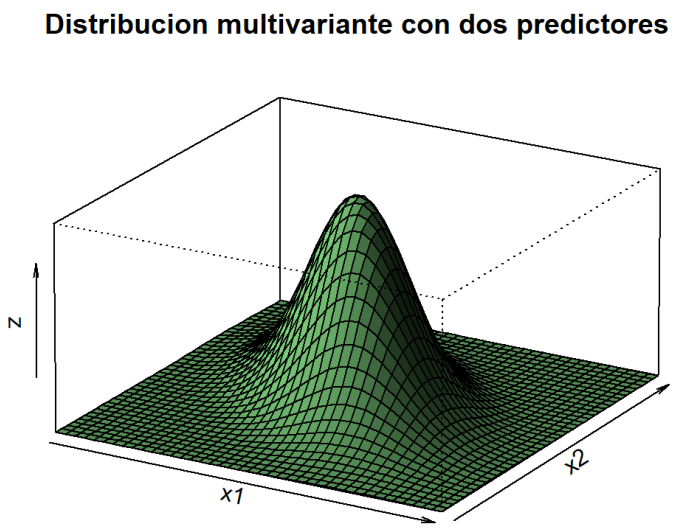


Figura .. representaciones gráficas de distribuciones normales multivariante

```
# Otra forma de representar una distribucion bivariante
library(mvtnorm)
library(scatterplot3d)
sigma.zero <- matrix(c(1,0,0,1), ncol=2)
x1000      <- rmvnorm(n=1000, mean=c(0,0), sigma=sigma.zero)
scatterplot3d(x1000[,1], x1000[,2], dmvnorm(x1000, mean=c(0,0), sigma=sigma.zero),
              highlight=TRUE, xlab = "x", ylab = "y", zlab = "z", )
```

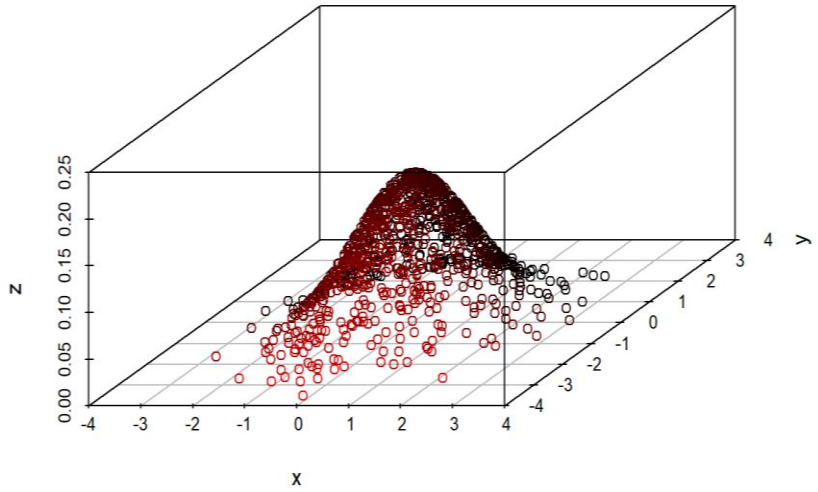


Figura. representaciones gráficas de distribuciones normales multivariante

Para indicar que una variable aleatoria p-dimensional X sigue una distribución normal multivariante se emplea la terminología $X \sim (\mu, \sum)$. Donde μ es el vector promedio de X y $\sum$ es la covarianza de X, que al ser un vector con p elementos, es una matriz p x p con la covarianza de cada par de predictores. La ecuación que define la función de densidad de una distribución normal multivariante es:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}}\ exp\ (-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu))$$

Si se sigue el mismo procedimiento que el mostrado para LDA con un solo predictor, pero esta vez con la ecuación de multivariante normal, y se asume que la matriz de covarianzas $\sum$ es igual para las K clases, se obtiene que el clasificador de Bayes es:

$$\hat{\delta}_k\ (x) = log(P(Y = k|X = x)) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + log(\pi_k)$$

Cuando los parámetros poblacionales se desconocen, no se puede calcular el límite de decisión de Bayes exacto, por lo que se recurre a la estimación de $\mu_1 \ldots, \mu_k,\ \pi_1 \ldots \pi_k$ y $\sum$ para obtener los límites de decisión de LDA.

11. Precisión del LDA

Una vez que las normas de clasificación se han establecido, se tiene que evaluar como de buena es la clasificación resultante. En otras palabras, evaluar el porcentaje de aciertos en las clasificaciones.

Las matrices de confusión son una de las mejores formas de evaluar la capacidad de acierto que tiene un modelo LDA. Muestran el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. El método LDA busca los límites de decisión que más se aproximan al clasificador de Bayes, que, por definición, tiene la menor ratio de error total de entre todos los clasificadores (si se cumple la condición de normalidad). Por lo tanto, el LDA intenta conseguir el menor número de clasificaciones erróneas posibles, pero no diferencia entre falsos positivos o falsos negativos. Si se quiere intentar reducir el número de errores de clasificación en una dirección determinada (por ejemplo, menos falsos

negativos) se puede modificar el límite de decisión, aunque como consecuencia aumentará el número de falsos positivos.

Cuando para evaluar el error de clasificación se emplean las mismas observaciones con las que se ha creado el modelo, se obtiene lo que se denomina el training error. Si bien esta es una forma sencilla de estimar la precisión en la clasificación, tiende a ser excesivamente optimista. Es más adecuado evaluar el modelo empleando observaciones nuevas que el modelo no ha visto, obteniendo así el test error.

Algoritmo del LDA.

1) Aplicar escalado de variables a la matriz de características X, compuesta por n variables independientes
2) Si C es el numero de clases, calculamos C vectores m-dimensionales, de modo que cada uno contenga las medias de las características de las observaciones para cada clase.
   Ejemplo: Supongamos que las variables dependientes tienen dos clases, codificadas como 0 y 1, y sea X_j^i la característica j-esima de la observación i-esima,

$$\boldsymbol{\mu_0} = \begin{vmatrix} \frac{1}{n_0} \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 0}} x_1^i \\ \vdots \\ \frac{1}{n_0} \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 0}} x_m^i \end{vmatrix}$$

   Medias de todas las columnas de la clase 0

$$\boldsymbol{\mu_1} = \begin{vmatrix} \frac{1}{n_1} \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 1}} x_1^i \\ \vdots \\ \frac{1}{n_1} \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 1}} x_m^i \end{vmatrix}$$

   Medias de todas las columnas de la clase 1

   Esto genera dos vectores de medias.

3) Calculamos la matriz de productos cruzados centrados en la media para cada clase, que mide la varianza dentro de cada clase.
   Ejemplo: para dos clases 0 y 1, las dos matrices de productos cruzados So y S1 para las respectivas clases 0 y 1 son:

$$S_0 = \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 0}} \left((x_1^i, \dots, x_m^i) - \boldsymbol{\mu_0}\right) \left((x_1^i, \dots, x_m^i) - \boldsymbol{\mu_0}\right)^T$$

$$S_1 = \sum_{\substack{i=1,\dots,n \\ y^i \in \text{class } 1}} \left((x_1^i, \dots, x_m^i) - \boldsymbol{\mu_1}\right) \left((x_1^i, \dots, x_m^i) - \boldsymbol{\mu_1}\right)^T$$

4) Calculamos la covarianza normalizada de todas las matrices anteriores, W. Ejemplo: con dos clases 0 y 1, la covarianza normalizada W es:

$$W = \frac{1}{n_0} S_0 + \frac{1}{n_1} S_1$$

5) Calculamos la matriz de covarianzas global entre clases, B

   Para el ejemplo, la matriz de covarianza global entre clases, es:

$$B = n_0(\boldsymbol{\mu_0} - \boldsymbol{\mu})(\boldsymbol{\mu_0} - \boldsymbol{\mu})^T + n_1(\boldsymbol{\mu_1} - \boldsymbol{\mu})(\boldsymbol{\mu_1} - \boldsymbol{\mu})^T$$

$$\text{where } \boldsymbol{\mu} = \underbrace{(\mu, \dots, \mu)}_{m \text{ times}}^T \text{ with } \mu = \frac{1}{n} \sum_{i=1,\dots,n} \sum_{j=1,\dots,m} x_j^i$$

6) Calculamos los valores y vectores propios de la matriz $W^{-1}B$
7) Elegimos los p valores propios mas grandes como el número de dimensiones reducidas.
8) Los p vectores propios asociados a los p valores propios mas grandes son los discriminantes lineales. El especio m-dimensional del dataset original se proyecta al nuevo subespacio p-dimensional de características, aplicando la matriz de proyecciones (que tiene los p vectores propios por columnas)

12. Variables cualitativas en el análisis discriminante.

Si entre las variables independientes encuentra una o más variables cualitativas, sus valores deben ser recodificadas a valores numéricos que correspondan en algún sentido a las variables originales.

En variables con dos categorías: se puede recodificar como ***0 y 1***, el valor 1 representa la presencia de la cualidad y el 0 la ausencia (en todo caso la presencia de la otra).

En variables con más de dos categorías: deberán generarse tantas variables como el total menos 1 categorías.

Ejemplo: Si los valores en la variable TABACO son 1 y 2, y el valor 2 representa a la variable ha dejado de fumar. 2 se puede recodificar como 0 y 1, fumador.

13. Ejemplo datos insectos en R

Un equipo de biólogos quiere generar un modelo estadístico que permita identificar a que especie (a o b) pertenece un determinado insecto. Para ello se han medido tres variables (longitud de las patas, diámetro del abdomen y diámetro del órgano sexual) en 10 individuos de cada una de las dos especies.

Datos de entrenamiento (existen varias formas de ingreso de datos a R, puede ser de base de datos, Excel, csv, vectores, data frames, etc, veamos la función input)

```
# Obtencion de los datos de entrenamiento
input <- ("
especie pata abdomen organo_sexual
a 191 131 53
a 185 134 50
a 200 137 52
a 173 127 50
a 171 128 49
a 160 118 47
a 188 134 54
a 186 129 51
a 174 131 52
a 163 115 47
b 186 107 49
b 211 122 49
b 201 144 47
b 242 131 54
b 184 108 43
b 211 118 51
b 217 122 49
b 223 127 51
b 208 125 50
b 199 124 46 ")
datos <- read.table(textConnection(input), header = TRUE)
head(datos)     # muestra los primeros 6 datos de cada variable
```

```
##   especie pata abdomen organo_sexual
## 1       a  191     131           53
## 2       a  185     134           50
## 3       a  200     137           52
## 4       a  173     127           50
## 5       a  171     128           49
## 6       a  160     118           47
```

```
str(datos) # muestra la estructura de las variables
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ especie      : chr  "a" "a" "a" "a" ...
##  $ pata         : int  191 185 200 173 171 160 188 186 174 163 ...
##  $ abdomen      : int  131 134 137 127 128 118 134 129 131 115 ...
##  $ organo_sexual: int  53 50 52 50 49 47 54 51 52 47 ...
```

observe:

1) indica que se tiene 4 variables cada una contiene 20 observaciones
2) la variable especie es de tipo chr (cualitativa con dimensiones a y b), es la variable dependiente y el LDA necesita ser de tipo factor.
3) las otras variables son de tipo int (cuantitativos). es recomendable que las variables independientes sean cuantitativas por lo que se puede realizar pruebas de normalidad, pero se puede incluir variables cualitativas, pero codificadas por ejemplo como variables dummy. Caso contrario usar la logística.

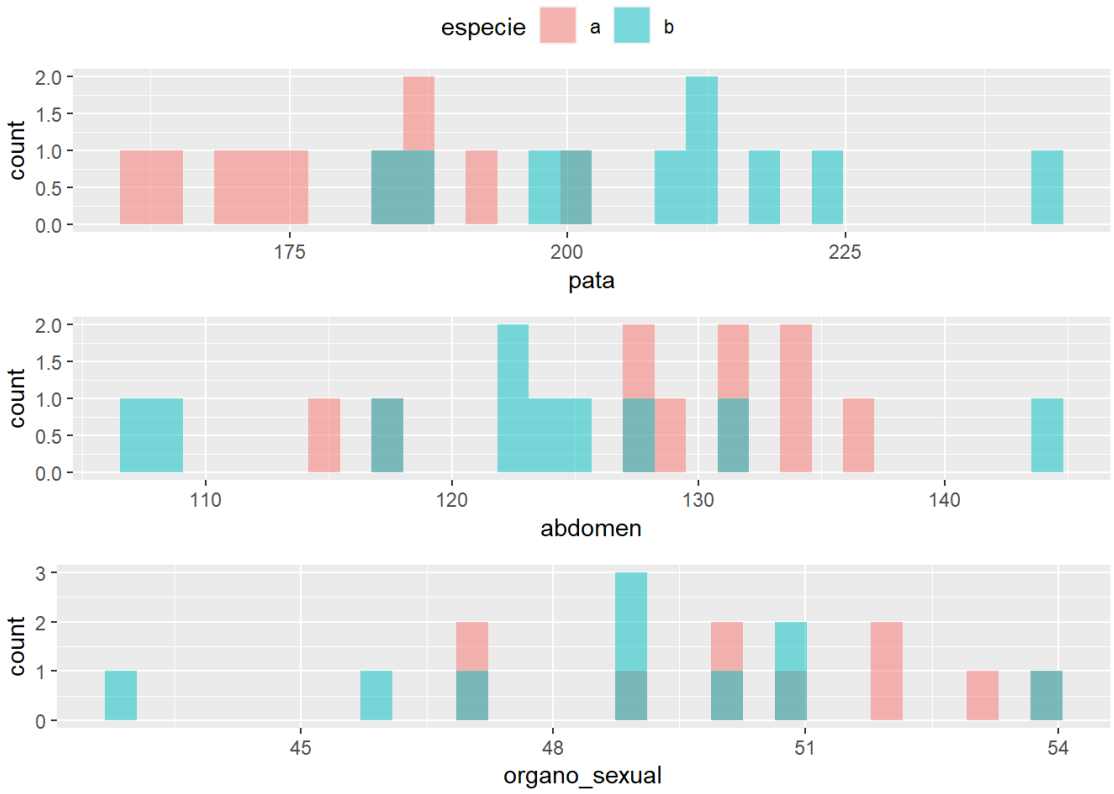
```
datos$especie <- as.factor(datos$especie)
str(datos)
```

```
## 'data.frame':    20 obs. of  4 variables:
##  $ especie      : Factor w/ 2 levels "a","b": 1 1 1 1 1 1 1 1 1 1 ...
##  $ pata         : int  191 185 200 173 171 160 188 186 174 163 ...
##  $ abdomen      : int  131 134 137 127 128 118 134 129 131 115 ...
##  $ organo_sexual: int  53 50 52 50 49 47 54 51 52 47 ...
```

Exploración gráfica.

no se olvide de instalar librerías antes de llamarlas

```
# Exploración grafica de los datos
library(ggplot2)  # llamando a la librería ggplot2
library(gridExtra)
library(ggpubr)
p1 <- ggplot(data = datos, aes(x = pata, fill = especie)) +
  geom_histogram(position = "identity", alpha = 0.5)
p2 <- ggplot(data = datos, aes(x = abdomen, fill = especie)) +
  geom_histogram(position = "identity", alpha = 0.5)
p3 <- ggplot(data = datos, aes(x = organo_sexual, fill = especie)) +
  geom_histogram(position = "identity", alpha = 0.5)
ggarrange(p1, p2, p3, nrow = 3, common.legend = TRUE)
```



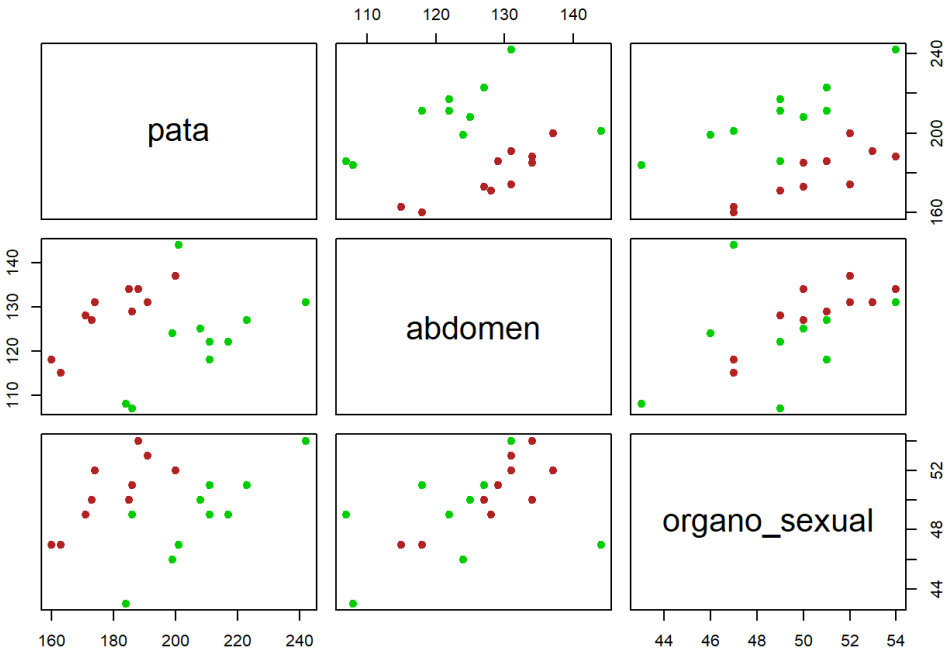
En el gráfico de puede observar el solapamiento de cada variable:

1) La variable longitud de pata indica menor solapamiento debido a que no existe muchas barras mezcladas, además de que los grupos están casi agrupados (azules por un lado y rojos por el otro)
2) Observe que las variables abdomen y órgano sexual el solapamiento es mayor, es decir, vemos que las barras se mezclan mucho.

A nivel individual, la longitud de la pata parece ser la variable que más se diferencia entre especies (menor solapamiento entre poblaciones).

Gráfico de dispersión entre variables

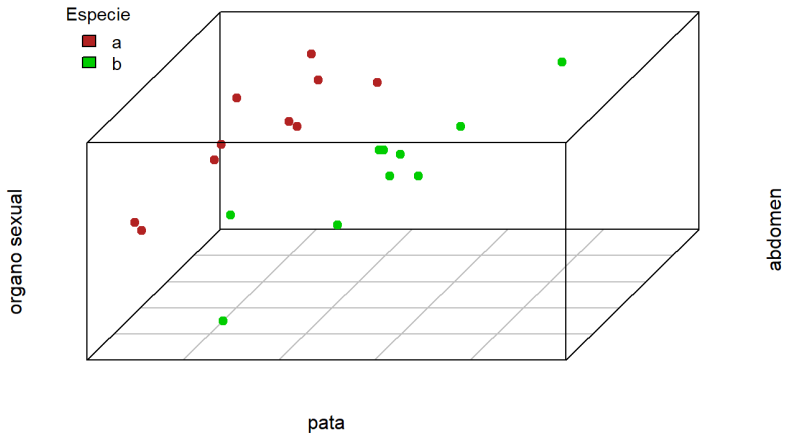
```
# correlaciones
pairs(x = datos[, c("pata", "abdomen", "organo_sexual")],
      col = c("firebrick", "green3")[datos$especie], pch = 19)
```



En primer lugar, el par de variables abdomen y pata parecen separar bien las dos especies y en segundo logar el par de variables pata y organo_sexual (con algunos errores se puede trazar una línea de separación).

Por tener tres variables independientes en estudio, podemos recurrir a un gráfico en 3d.

```
library(scatterplot3d)
scatterplot3d(datos$pata, datos$abdomen, datos$organo_sexual,
              color = c("firebrick", "green3")[datos$especie],
              pch = 19, grid = TRUE, tick.marks = FALSE, xlab = "pata",
              ylab = "abdomen", zlab = "organo sexual", angle = 65)
legend("topleft", bty = "n", cex = .9, title = "Especie", c("a", "b"),
       fill = c("firebrick", "green3"))
```



La representación de las tres variables de forma simultánea parece indicar que las dos especies (a y b) sí están bastante separadas en el espacio 3D (para tres variables).

Prior probabilities. Probabilidades previas

Como no se dispone de información sobre la abundancia relativa de las especies a nivel poblacional, se considera como probabilidad previa de cada especie el número de observaciones de la especie entre el número de observaciones totales.

$$\hat{\pi}_a = \hat{\pi}_b = \frac{10}{20} = 0.5$$

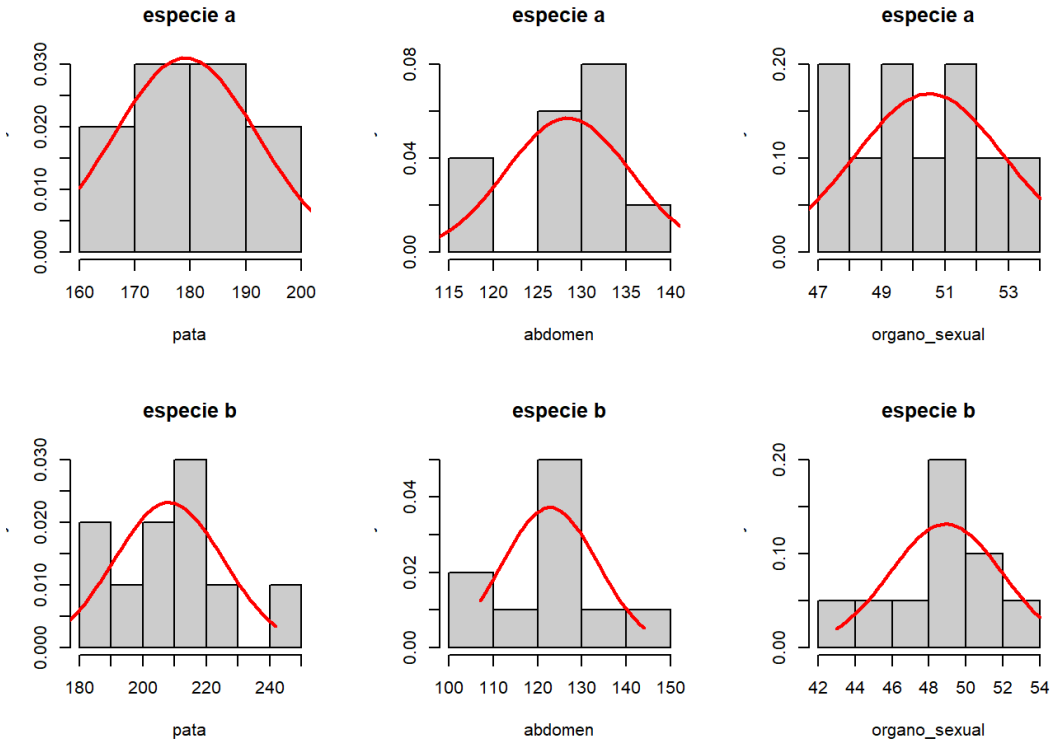
Prueba de normalidad

Existen varias formas de observar la normalidad, en este caso puede observarse la normalidad para cada variable o utilizar la normalidad multivariante.

Observando histogramas:

```
# Distribución de los predictores de forma individual
# Representación mediante histograma de cada variable para cada especie
par(mfcol = c(2, 3))  # solicito que los gráficos estén ordenados en 2 filas y 3 columnas
for (k in 2:4) {
  j0 <- names(datos)[k]
  # br0 <- seq(min(datos[, k]), max(datos[, k]), le = 11)
  x0 <- seq(min(datos[, k]), max(datos[, k]), le = 50)
  for (i in 1:2) {
    i0 <- levels(datos$especie)[i]
    x <- datos[datos$especie == i0, j0]
    hist(x, proba = T, col = grey(0.8), main = paste("especie", i0), xlab = j0)
    lines(x0, dnorm(x0, mean(x), sd(x)), col = "red", lwd = 2)
  }
```

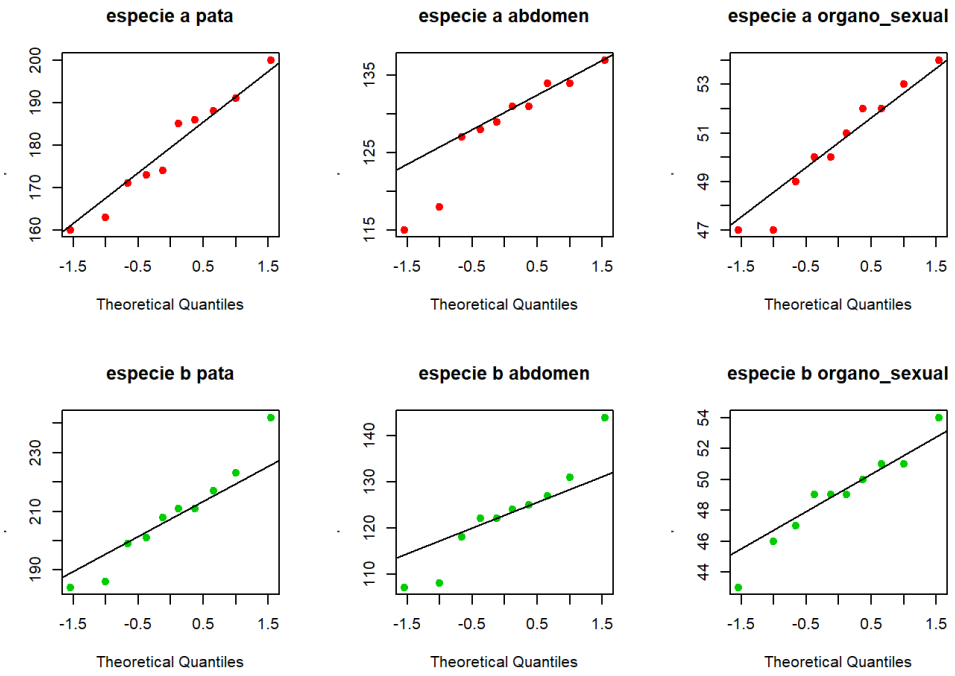
```
}
```



No se muestra claramente la normalidad de los datos excepto en especie a variables longitud de pata. recurramos al grafico de cuantiles.

```
# Representación de cuantiles normales de cada variable para cada especie
for (k in 2:4) {
 j0 <- names(datos)[k]
 x0 <- seq(min(datos[, k]), max(datos[, k]), le = 50)
 for (i in 1:2) {
  i0 <- levels(datos$especie)[i]
  x <- datos[datos$especie == i0, j0]
  qqnorm(x, main = paste("especie", i0, j0), pch = 19, col = i + 1)
  qqline(x)
 }
 }
```

Igualmente puede usar librerías para mejorar los gráficos



Igualmente, tenemos sospecha de que las variables abdomen para a y b no tenderían a la normal. Finalmente, utilicemos pruebas de Hipótesis, en este caso la más adecuada es la Shapiro por tratarse de datos pequeños. (desde luego existen otras pruebas).

```
# Contraste de normalidad Shapiro-Wilk para cada variable en cada especie
library(reshape2)
library(knitr)
library(dplyr)
datos_tidy <- melt(datos, value.name = "valor")
kable(datos_tidy %>% group_by(especie, variable) %>%
      summarise(p_value_Shapiro.test = shapiro.test(valor)$p.value))
```

especie	variable	p_value_Shapiro.test
a	pata	0.7763034
a	abdomen	0.1845349
a	organo_sexual	0.6430844
b	pata	0.7985711
b	abdomen	0.5538213
b	organo_sexual	0.8217855

No hay evidencias de falta de normalidad univariante en ninguna de las variables empleadas como predictores en ninguno de los grupos. (en todos los casos $p > 0.05$)

Normalidad multivariante:

El paquete MVN contiene funciones que permiten realizar los tres test de hipótesis comúnmente empleados para evaluar la normalidad multivariante (Mardia, Henze-Zirkler y Royston) y también funciones para identificar outliers que puedan influenciar en el contraste.

```
# contrastes de normalidad multivariante
library(MVN)
# test de normalidad multivariante de Mardia
result <- mvn(data = datos[,-1], mvnTest = "mardia")
result$multivariateNormality
```

```
##              Test         Statistic          p value Result
## 1 Mardia Skewness  8.69011058284102 0.561743937637725    YES
## 2 Mardia Kurtosis -0.088170074036502 0.929741502120227    YES
## 3             MVN              <NA>              <NA>    YES
```

```
# test de normalidad multivariante de Henze-Zirkler's
result <- mvn(data = datos[,-1], mvnTest = "hz")
result$multivariateNormality
```

```
##           Test           HZ         p value     MVN
## 1 Henze-Zirkler 0.7870498 0.07666139 YES
```

```
# # test de normalidad multivariante de Royston's
result <- mvn(data = datos[,-1], mvnTest = "royston")
result$multivariateNormality
```

```
##     Test         H          p value    MVN
## 1 Royston 0.4636176 0.9299447 YES
```

Los tres test indican que existe normalidad multivariante.

Homogeneidad de Varianza

De entre los diferentes test que contrastan la homogeneidad de varianza, el más recomendable cuando solo hay un predictor, dado que se asume que se distribuye de forma normal, es el test de *Barttlet*. Cuando se emplean múltiples predictores, se tiene que contrastar que la matriz de covarianzas ($\sum$) es constante en todos los grupos, siendo recomendable comprobar también la homogeneidad de varianza para cada predictor a nivel individual.

El test M Box desarrollado por el matemático Box (1949) como una extensión del test de Barttlet para escenarios multivariante y permite contrastar la igualdad de matrices entre grupos. El test Box M es muy sensible a violaciones de la normalidad multivariante, por lo que esta debe ser contrastada con anterioridad. Ocurre con frecuencia, que el resultado de un test M Box resulta significativo debido a la falta de distribución normal multivariante en lugar de falta de homogeneidad en las matrices de covarianza. Dada la sensibilidad de este test se recomienda emplear un límite de significancia de 0.001 (Tabachnick & Fidell, 2001, y http://www.real-statistics.com/multivariate-statistics/).

```
# contraste de matrices de covarianza
library(rpanel)
library(tcltk)
library(tkrplot)
library(MASS)
library(biotools)
boxM(data = datos[, 2:4], grouping = datos[, 1])
```

```
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  datos[, 2:4]
## Chi-Sq (approx.) = 9.831, df = 6, p-value = 0.132
```

Se puede aceptar que la matriz de covarianza es igual en todos los grupos $p(0.132) > \alpha(0.05)$.

Estimación de los parámetros de la función de densidad ($\hat{\mu}(X), \sum$) y cálculo de la función discriminante.

Estos dos pasos se realizan mediante la función lda( ) del paquete MASS. lda( ) realiza la clasificación mediante la aproximación de Fisher.

```
# Estimación de los parámetros  y cálculo de la función discriminante
```

```
modelo_lda <- lda(formula = especie ~ pata + abdomen + organo_sexual, data = datos)
modelo_lda
```

```
## Call:
## lda(especie ~ pata + abdomen + organo_sexual, data = datos)
##
## Prior probabilities of groups:
##   a   b
## 0.5 0.5
##
## Group means:
##    pata  abdomen organo_sexual
## a 179.1   128.4         50.5
## b 208.2   122.8         48.9
##
## Coefficients of linear discriminants:
##                              LD1
## pata                  0.13225339
## abdomen           -0.07941509
## organo_sexual   -0.52655608
```

Función discriminante:

especimen = 0.13225330(pata)-0.079441509(abdomen) – 0.52655608(organo_sexual)

obteniendo la predicción para cada dato.

```
# evaluación de los errores de clasificación
# clase predicha
clasif<-predict(modelo_lda, datos[-1])$class
clasif
```

```
#  [1] a a a a a a a a a a b b b b b b b b b b
## Levels: a b
```

obtenemos las probabilidades de las predicciones

```
# probabilidad predecida
predicciones <- predict(modelo_lda, datos[,-1],)$posterior
predicciones
```

```
##              a           b
## 1  9.999976e-01 2.432688e-06
## 2  9.999594e-01 4.056357e-05
## 3  9.985810e-01 1.419037e-03
## 4  9.999998e-01 2.033594e-07
## 5  9.999995e-01 5.195408e-07
## 6  9.999961e-01 3.899541e-06
## 7  1.000000e+00 6.241502e-09
## 8  9.999589e-01 4.114531e-05
## 9  1.000000e+00 3.514118e-10
## 10 9.998983e-01 1.016987e-04
## 11 1.361196e-02 9.863880e-01
## 12 2.643877e-07 9.999997e-01
## 13 8.250009e-03 9.917500e-01
## 14 5.548701e-09 1.000000e+00
## 15 7.248486e-09 1.000000e+00
## 16 1.155103e-05 9.999884e-01
## 17 4.490952e-09 1.000000e+00
## 18 1.309176e-07 9.999999e-01
## 19 1.030374e-04 9.998970e-01
## 20 6.207930e-07 9.999994e-01
```

```
str(predicciones)
```

```
##  num [1:20, 1:2] 1 1 0.999 1 1 ...
##  - attr(*, "dimnames")=List of 2
```

```
##   ..$ : chr [1:20] "1" "2" "3" "4" ...
##   ..$ : chr [1:2] "a" "b"
```

nos interesa la segunda probabilidad

```
predicciones[,2]
```

```
##           1           2           3           4           5           6
## 2.432688e-06 4.056357e-05 1.419037e-03 2.033594e-07 5.195408e-07 3.899541e-06
##           7           8           9          10          11          12
## 6.241502e-09 4.114531e-05 3.514118e-10 1.016987e-04 9.863880e-01 9.999997e-01
##          13          14          15          16          17          18
## 9.917500e-01 1.000000e+00 1.000000e+00 9.999884e-01 1.000000e+00 9.999999e-01
##          19          20
## 9.998970e-01 9.999994e-01
```

Almacenamos las predicciones conjuntamente con la data original

```
# Almacenamiento de datos con clase y probabilidad predicha
datosI=cbind(datos,clasif,predicciones[,2])
datosI
```

```
##    especie pata abdomen organo_sexual clasif predicciones[, 2]
## 1        a  191    131            53       a      2.432688e-06
## 2        a  185    134            50       a      4.056357e-05
## 3        a  200    137            52       a      1.419037e-03
## 4        a  173    127            50       a      2.033594e-07
## 5        a  171    128            49       a      5.195408e-07
## 6        a  160    118            47       a      3.899541e-06
## 7        a  188    134            54       a      6.241502e-09
## 8        a  186    129            51       a      4.114531e-05
## 9        a  174    131            52       a      3.514118e-10
## 10       a  163    115            47       a      1.016987e-04
## 11       b  186    107            49       b      9.863880e-01
## 12       b  211    122            49       b      9.999997e-01
```

```
## 13     b  201    144        47          b     9.917500e-01
## 14     b  242    131        54          b     1.000000e+00
## 15     b  184    108        43          b     1.000000e+00
## 16     b  211    118        51          b     9.999884e-01
## 17     b  217    122        49          b     1.000000e+00
## 18     b  223    127        51          b     9.999999e-01
## 19     b  208    125        50          b     9.998970e-01
## 20     b  199    124        46          b     9.999994e-01
```

en clasif esta los valores predichos

guardamos estos datos en un archivo tabla con val. predichos.csv. podemos recuperarlo en Excel como texto.

```
write.csv(datosI,"tabla con val.predichos.csv")
```

Evaluación de los errores de clasificación.

```
# evaluación de los errores de clasificación (Matriz de confusión)
addmargins(table(datos$especie,clasif))
```

```
##             clasif
##              a     b   Sum
##   a         10    0    10
##   b          0    10   10
##   Sum 10   10   20
```

La table muestra la bondad del modelo (algo así como el coeficiente de determinación)
10: no hubo cambios,
0: representan los cambios de a a b
0: representan los cambio de b a a
10: representan los cambios de b a b

El modelo es perfecto, debido a que no existe cambios
representando como probabilidades

```
prop.table(table(datos$especie,clasif),1)
```

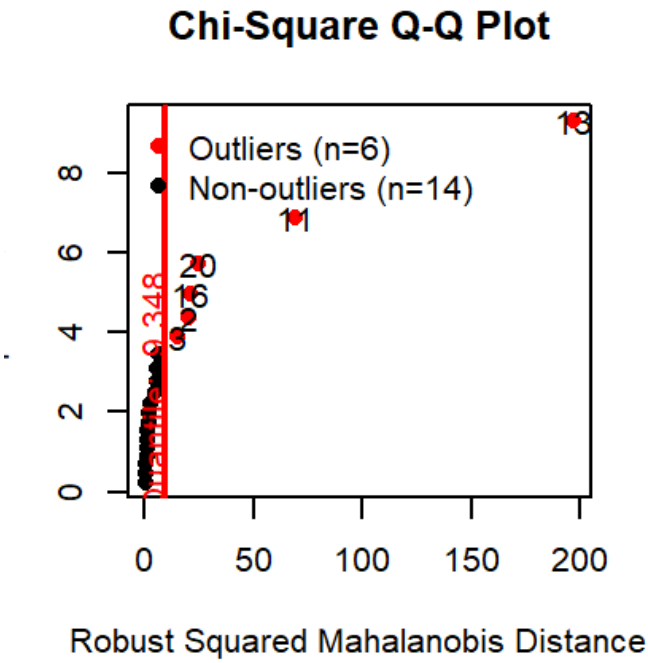
```
##    clasif
##       a  b
##  a    1  0
##  b    0  1
```

```
# Tasa de Clasificación
er=mean(predicciones!= datos$especie)
1-err
```

```
## [1] 0
```

Observando Outtliers

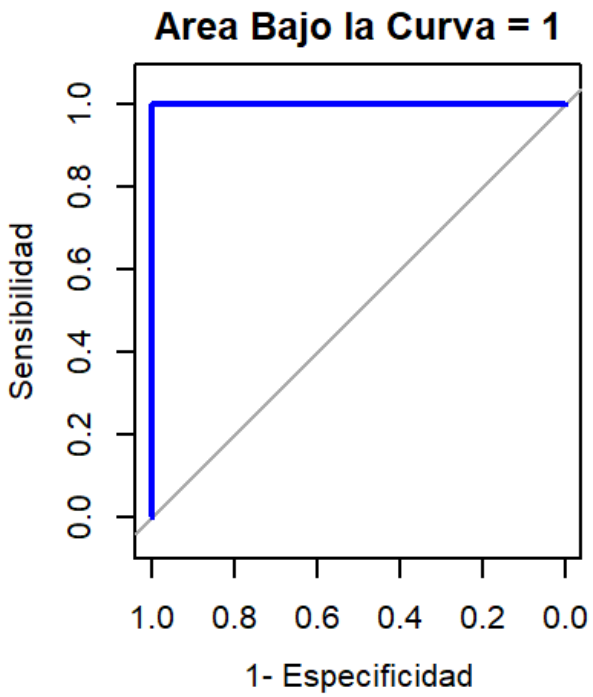
```
par(mfcol = c(1, 1))
```



Curva ROC: la curva es usada para verificar la bondad del modelo.

```
# Curva ROC y Área bajo la curva
library(pROC)
```

```
datos$especie=as.numeric(datos$especie)
clasif=as.numeric(clasif)
clasif=predicciones[,2]
#Calculando el area bajo la curva
areaROC<-auc(roc(datos$especie,clasif))
areaROC
ROC1<-plot.roc(datos$especie,clasif, xlab="1- Especificidad", ylab="Sensibilidad",
        main = paste('Area Bajo la Curva =',round(areaROC,3)), col="blue")
```



Como vemos la línea azul está alejada de la diagonal, es decir el modelo es bueno.

Si la línea azul estaría cerca de la diagonal, ese modelo no es bueno.

```
error <- mean(datos$especie!= predicciones$class) * 100
paste("error=", error, "%")
```

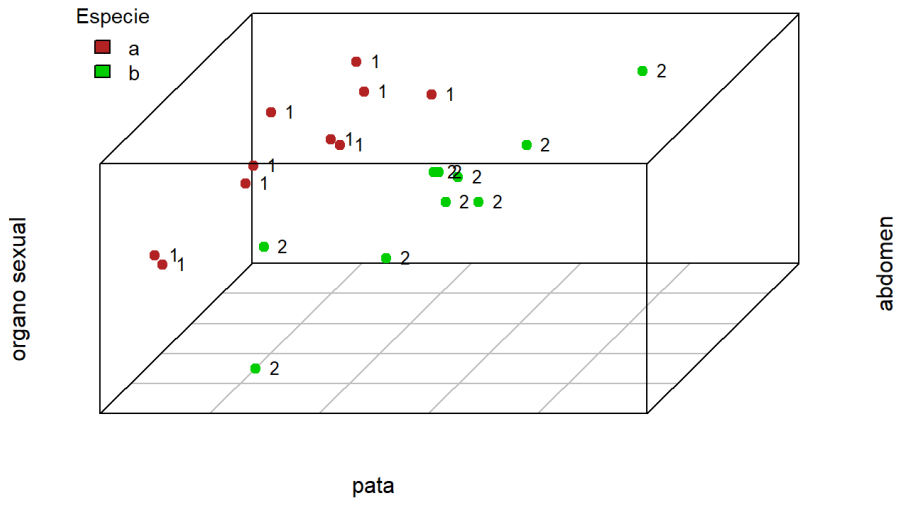
```
## [1] "error= 0 %"
```

Empleando las mismas observaciones con las que se ha generado el modelo discriminante, la precisión de clasificación es del 100%. Evaluar un modelo con los mismos datos con los que se ha creado suele resultar estimaciones de la precisión demasiado optimistas. La

estimación del test error mediante validación cruzada es más adecuada para obtener una evolución realista del modelo.

La siguiente imagen muestra la representación de las observaciones, coloreadas por la verdadera especie a la que pertenecen y acompañadas por una etiqueta con la especie que ha predicho el LDA.

```
# validacion
library(scatterplot3d)
with(datos, {
  s3d <- scatterplot3d(pata, abdomen, organo_sexual,
    color = c("firebrick", "green3")[datos$especie], pch = 19,
    grid = TRUE, tick.marks = FALSE, xlab = "pata",
    ylab = "abdomen", zlab = "organo sexual", angle = 65)
  s3d.coords <- s3d$xyz.convert(pata, abdomen, organo_sexual)
  # convierte coordenadas 3D en proyecciones 2D
  text(s3d.coords$x, s3d.coords$y,
       # cordenadas x, y
       labels = datos$especie,
       # texto
       cex = .8, pos = 4)
  legend("topleft", bty = "n", cex = .9, title = "Especie",
         c("a", "b"), fill = c("firebrick", "green3"))
  })
```



Predicción:

Una vez obtenidas las funciones discriminantes, se puede clasificar un nuevo insecto en función de sus medidas.
Por ejemplo, un nuevo espécimen cuyas medidas sean: pata = 194, abdomen = 124, organo_sexual = 49.

```
# predict(object = modelo_lda, datos)
nuevas_observaciones <- data.frame(pata = 194, abdomen = 124, organo_sexual = 49)
nuevas_observaciones
```

```
##   pata abdomen organo_sexual
## 1  194     124           49
```

```
predict(object = modelo_lda, newdata = nuevas_observaciones)
```

```
## $class
## [1] b
## Levels: a b
## $posterior
##              a             b
## 1   0.05823333     0.9417667
## $x
##        LD1
## 1 0.5419421
```

El resultado muestra que, según la función discriminante, la probabilidad posterior de que el espécimen pertenezca a la especie b es del 94.2% frente al 5.8% de que pertenezca a la especie a.

14. Ejemplo con Iris data

El set de datos Iris contiene métricas de 150 flores de 3 especies diferentes de planta Iris. Para cada flor se han registrado 4 variables: sepal length, sepal width, petal length y petal width, todas ellas en centímetros. Se desea generar un modelo discriminante que permita clasificar las flores en las distintas especies empleando las variables mencionadas.

```
# ejemplo IRIS data
```

```
# ejemplo de reclasificacion
data("iris")
```

```
head(iris)
```

```
##   Sepal.Length  Sepal.Width  Petal.Length Petal.Width 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
```

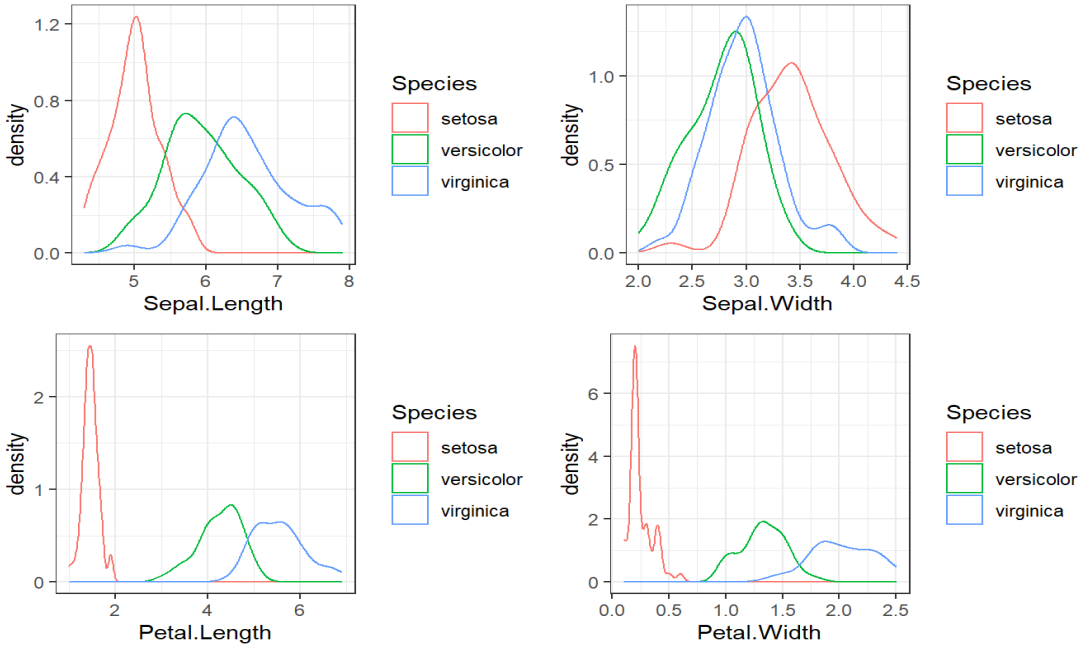
```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ..
```

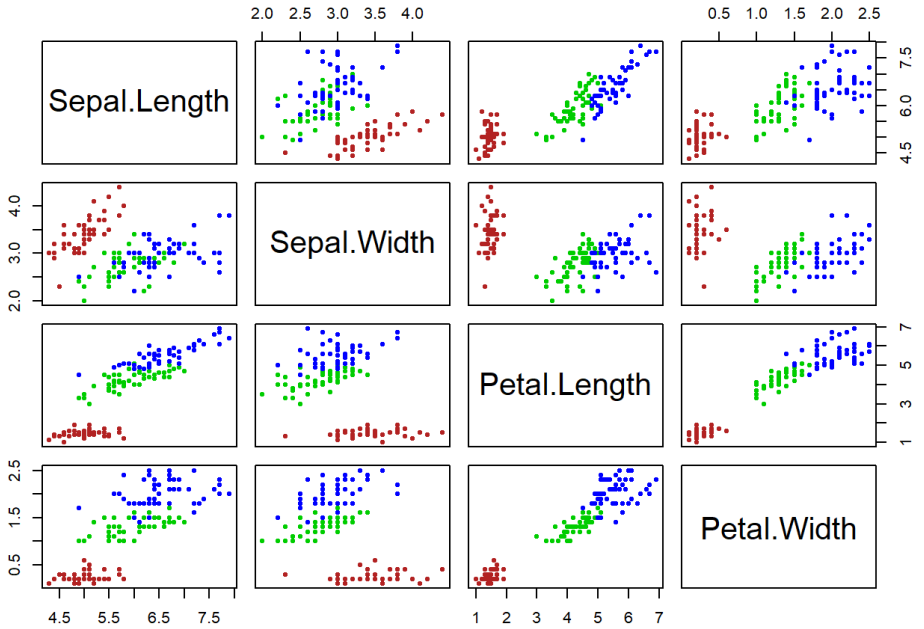
Exploración gráfica de los datos

```
# Exploracion grafica de los datos
library(ggplot2)
library(gridExtra)
p1 <- ggplot(data = iris, aes(x = Sepal.Length)) +
  geom_density(aes(colour=Species))+
  theme_bw()
p2 <- ggplot(data = iris, aes(x = Sepal.Width)) +
  geom_density(aes(colour=Species))+
  theme_bw()
p3 <- ggplot(data = iris, aes(x = Petal.Length)) +
  geom_density(aes(colour=Species))+
```

```
  theme_bw()
p4 <- ggplot(data = iris, aes(x = Petal.Width)) +
  geom_density(aes(colour=Species))+
  theme_bw()
grid.arrange(p1, p2, p3, p4)
```




```
# correlaciones
pairs(x = iris[,-5], col = c("firebrick", "green3", "blue")[iris$Species],
      pch = 20)
```



Las variables Petal.Lenght y Petal.Width son las dos variables con más potencial para poder separar entre clases. Sin embargo, están altamente correlacionadas, por lo que la información que aportan es en gran medida redundante.

Prior probabilities

Como no se dispone de información sobre la abundancia relativa de las especies a nivel poblacional, se considera como probabilidad previa de cada especie el número de observaciones de la especie entre el número de observaciones totales.

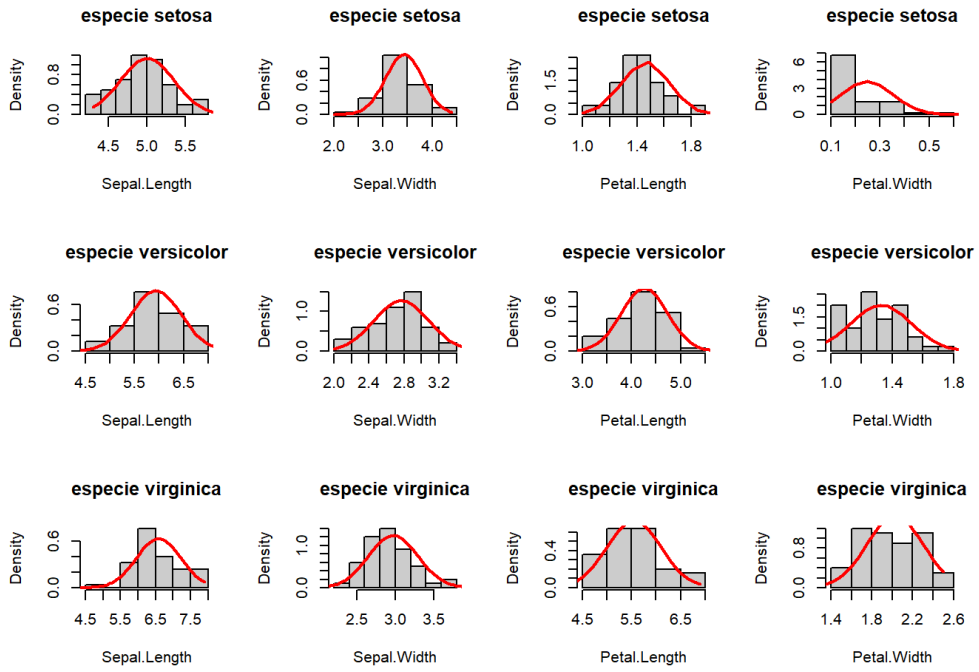
$$\hat{\pi}_{setosa} = \hat{\pi}_{versicolor} = \hat{\pi}_{virginica} = \frac{50}{150} = 0.33$$

Normalidad univariante, normalidad multivariante y homogeneidad de varianza

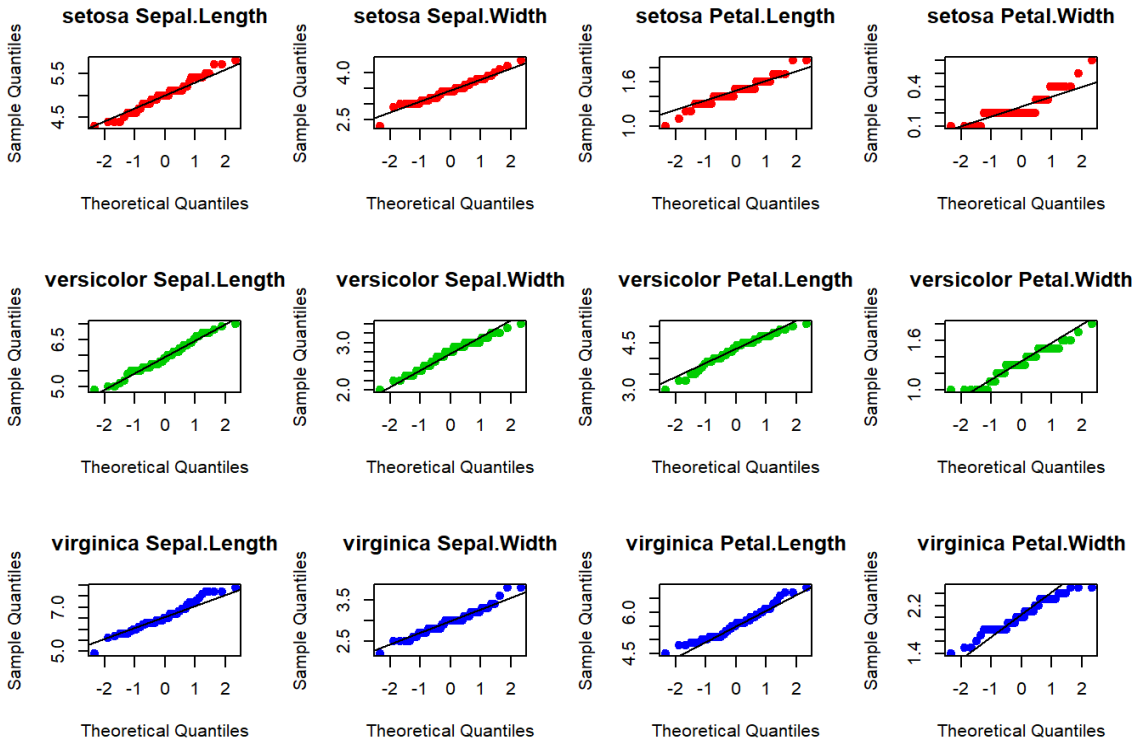
Distribución de los predictores de forma individual:

```
# Distribucion de los predictores de forma individual
# Representacion mediante histograma de cada variable para cada especie
X11()
par(mfcol = c(3, 4))
for (k in 1:4) {
  j0 <- names(iris)[k]
  x0 <- seq(min(iris[, k]), max(iris[, k]), le = 50)
  for (i in 1:3) {
    i0 <- levels(iris$Species)[i]
    x <- iris[iris$Species == i0, j0]
```

```
    hist(x, proba = T, col = grey(0.8), main = paste("especie", i0), xlab = j0)
    lines(x0, dnorm(x0, mean(x), sd(x)), col = "red", lwd = 2)
  }
}
```




```
# Representación de cuantiles normales de cada variable para cada especie
X11()
par(mfcol=c(3,4))
for (k in 1:4) {
  j0 <- names(iris)[k]
  x0 <- seq(min(iris[, k]), max(iris[, k]), le = 50)
  for (i in 1:3) {
    i0 <- levels(iris$Species)[i]
    x <- iris[iris$Species == i0, j0]
    qqnorm(x, main = paste(i0, j0), pch = 19, col = i + 1)
    qqline(x)
  }
}
```




```
# Contraste de normalidad Shapiro-Wilk para cada variable en cada especie
library(reshape2)
library(knitr)
library(dplyr)
datos_tidy <- melt(iris, value.name = "valor")
kable(datos_tidy %>% group_by(Species, variable) %>%
      summarise(p_value_Shapiro.test = round(shapiro.test(valor)$p.value,5)))
```

Species	variable	p_value_Shapiro.test
setosa	Sepal.Length	0.45951
setosa	Sepal.Width	0.27153
setosa	Petal.Length	0.05481
setosa	Petal.Width	0.00000
versicolor	Sepal.Length	0.46474
versicolor	Sepal.Width	0.33800
versicolor	Petal.Length	0.15848
versicolor	Petal.Width	0.02728
virginica	Sepal.Length	0.25831
virginica	Sepal.Width	0.18090
virginica	Petal.Length	0.10978
virginica	Petal.Width	0.08695

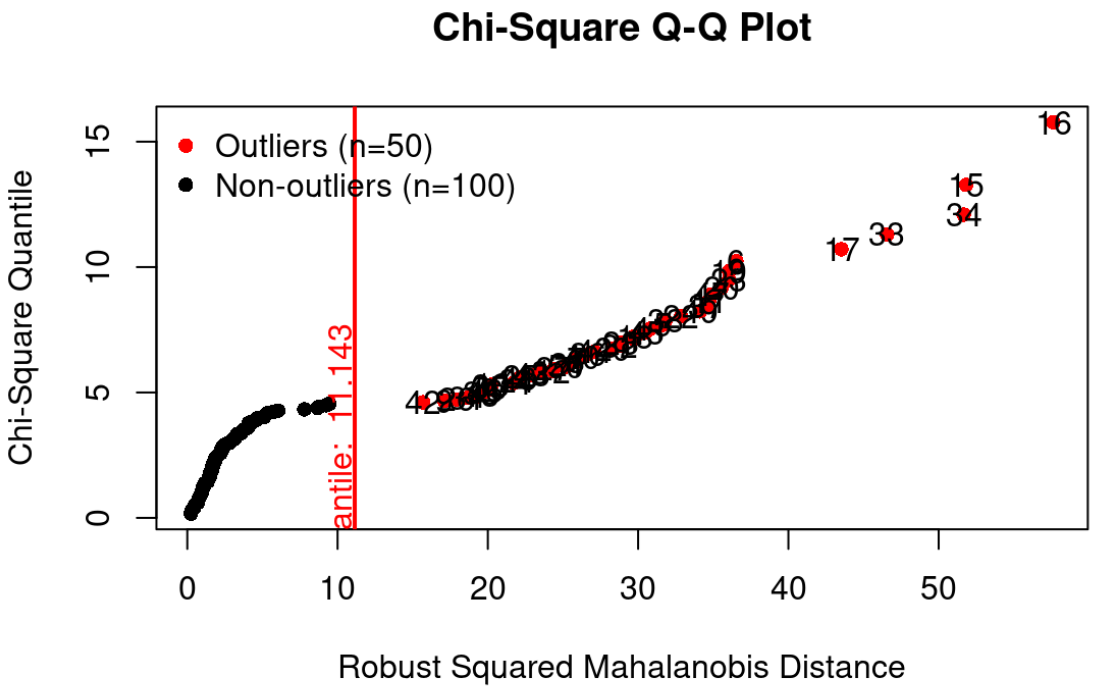
La variable *petal.width* no se distribuye de forma normal en los grupos setosa y versicolor.

Normalidad multivariante:

```
# normalidad multivariante
library(MVN)
# test de normalidad multivariante de Mardia
result <- mvn(data = iris[,-5], mvnTest = "mardia")
result$multivariateNormality
```

```
##              Test         Statistic             p value Result
## 1 Mardia Skewness  67.430508778062 4.75799820400869e-07     NO
## 2 Mardia Kurtosis -0.230112114481001   0.818004651478012    YES
## 3             MVN              <NA>              <NA>     NO
```

```
library(MVN)
outliers <- mvn(data = iris[,-5], mvnTest = "hz", multivariateOutlierMethod = "quan")
```




```
# test de normalidad multivariante de Henze-Zirkler's
```

```
result <- mvn(data = iris[,-5], mvnTest = "hz")
result$multivariateNormality
```

```
#           Test       HZ p value MVN
## 1 Henze-Zirkler 2.336394       0  NO
```

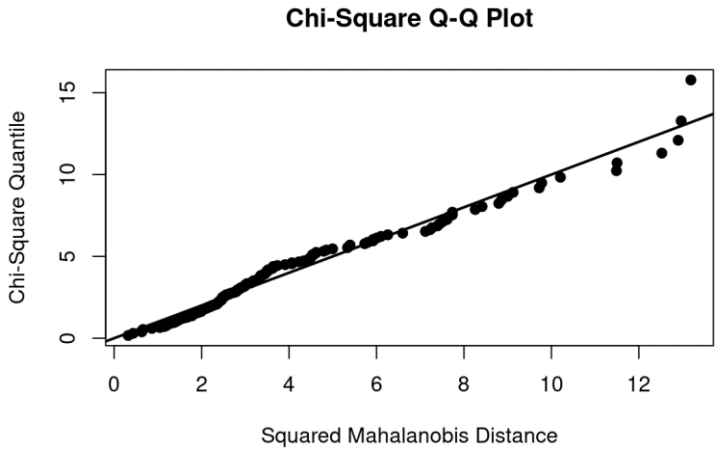
```
# test de normalidad multivariante de Royst
```

```
test de normalidad multivariante de Royston's
result <- mvn(data = iris[,-5], mvnTest = "royston")
result$multivariateNormality
```

```
##      Test       H      p value MVN
## 1 Royston 50.39667 3.098229e-11  NO
```

Los test muestran evidencias significativas de falta de normalidad multivariante. El LDA tiene cierta robustez frente a la falta de normalidad multivariante, pero es importante tenerlo en cuenta en la conclusión del análisis.

```
royston_test <- mvn(data = iris[,-5], mvnTest = "royston", multivariatePlot = "qq")
```




```
# contraste de matrices de covarianza
library(biotools)
```

```
boxM(data = iris[, -5], grouping = iris[, 5])
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  iris[, -5]
## Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

El test Box's M muestra evidencias de que la matriz de covarianza no es constante en todos los grupos, lo que a prioiri descartaría el método LDA en favor del QDA. Sin embargo, como el test Box's M es muy sensible a la falta de normalidad multivariante, con frecuencia resulta significativo no porque la matriz de covarianza no sea constante sino por la falta de normalidad, cosa que ocurre para los datos de Iris. Por esta razón se va a asumir que la matriz de covarianza sí es constante y que LDA puede alcanzar una buena precisión en la clasificación. En la evaluación del modelo se verá como de buena es esta aproximación. Además, en las conclusiones se debe explicar la asunción hecha.
Cálculo de la función discriminante

```
# calculo de la funcion discriminante
library(MASS)
modelo_lda <- lda(Species ~ Sepal.Width + Sepal.Length +
          Petal.Length + Petal.Width,data=iris)
modelo_lda
```

```
## Call:
## lda(Species ~ Sepal.Width + Sepal.Length + Petal.Length + Petal.Width,
##     data = iris)
##
## Prior probabilities of groups:
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Group means:
##            Sepal.Width Sepal.Length Petal.Length Petal.Width
## setosa           3.428        5.006        1.462       0.246
```

```
## versicolor       2.770       5.936       4.260       1.326
## virginica        2.974       6.588       5.552       2.026
##
## Coefficients of linear discriminants:
##                        LD1         LD2
## Sepal.Width   1.5344731  2.16452123
## Sepal.Length  0.8293776  0.02410215
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603  2.83918785
##
## Proportion of trace:
##    LD1    LD2
## 0.9912 0.0088
```

Modelo: 1.53447(Sepal.Width) + 0.82934(Sepal.Length) – 2.2012(Petal.Length)-2.8105(Petal.Width)

Evaluación de los errores de clasificación

```
# evaluacion de los errores de clasificacion
# clase predecida
clasific<-predict(modelo_lda, iris[-5])$class
head(clasific,20)
```

```
##  [1] setosa setosa setosa setosa setosa setosa
##  [7] setosa setosa setosa setosa setosa setosa
## [13] setosa setosa setosa setosa setosa setosa
## [19] setosa setosa
```

```
# probabilidad predecida
predic <- predict(modelo_lda, iris[,-5],)$posterior
head(predic,20)
```

```
##            setosa          versicolor       virginica
## 1   1.000000e+00 3.896358e-22 2.611168e-42
## 2   1.000000e+00 7.217970e-18 5.042143e-37
## 3   1.000000e+00 1.463849e-19 4.675932e-39
## 4   1.000000e+00 1.268536e-16 3.566610e-35
## 5   1.000000e+00 1.637387e-22 1.082605e-42
## 6   1.000000e+00 3.883282e-21 4.566540e-40
## 7   1.000000e+00 1.113469e-18 2.302608e-37
## 8   1.000000e+00 3.877586e-20 1.074496e-39
## 9   1.000000e+00 1.902813e-15 9.482936e-34
## 10  1.000000e+00 1.111803e-18 2.724060e-38
## 11  1.000000e+00 1.185277e-23 3.237084e-44
## 12  1.000000e+00 1.621649e-18 1.833201e-37
## 13  1.000000e+00 1.459225e-18 3.262506e-38
## 14  1.000000e+00 1.117219e-19 1.316642e-39
## 15  1.000000e+00 5.487399e-30 1.531265e-52
## 16  1.000000e+00 1.261505e-27 2.268705e-48
## 17  1.000000e+00 6.754338e-25 3.868271e-45
## 18  1.000000e+00 4.223741e-21 1.224313e-40
## 19  1.000000e+00 1.774911e-22 2.552153e-42
## 20  1.000000e+00 2.593237e-22 5.792079e-42
```

```
str(predic)
```

```
##  num [1:150, 1:3] 1 1 1 1 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:150] "1" "2" "3" "4" ...
##   ..$ : chr [1:3] "setosa" "versicolor" "virginica"
```

```
head(predic[,2], 20)
```

```
##            1            2            3            4            5            6
## 3.896358e-22 7.217970e-18 1.463849e-19 1.268536e-16 1.637387e-22 3.883282e-21
##            7            8            9           10           11           12
## 1.113469e-18 3.877586e-20 1.902813e-15 1.111803e-18 1.185277e-23 1.621649e-18
##           13           14           15           16           17           18
## 1.459225e-18 1.117219e-19 5.487399e-30 1.261505e-27 6.754338e-25 4.223741e-21
##           19           20
## 1.774911e-22 2.593237e-22
```

```
# Almacenamiento de datos con clase y probabilidad predecida
datosIr=cbind(iris,clasific,predic[,2])
head(datosIr,20)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species   clasific
## 1           5.1         3.5          1.4         0.2     setosa     setosa
## 2           4.9         3.0          1.4         0.2     setosa     setosa
## 3           4.7         3.2          1.3         0.2      setosa     setosa
## 4           4.6         3.1          1.5         0.2     setosa     setosa
## 5           5.0         3.6          1.4         0.2     setosa     setosa
## 6           5.4         3.9          1.7         0.4     setosa     setosa
## 7           4.6         3.4          1.4         0.3     setosa     setosa
## 8           5.0         3.4          1.5         0.2     setosa     setosa
## 9           4.4         2.9          1.4         0.2     setosa     setosa
## 10          4.9         3.1          1.5         0.1     setosa     setosa
## 11          5.4         3.7          1.5         0.2     setosa     setosa
## 12          4.8         3.4          1.6         0.2     setosa     setosa
## 13          4.8         3.0          1.4         0.1     setosa     setosa
## 14          4.3         3.0          1.1         0.1     setosa     setosa
## 15          5.8         4.0          1.2         0.2     setosa     setosa
## 16          5.7         4.4          1.5         0.4     setosa     setosa
## 17          5.4         3.9          1.3         0.4     setosa     setosa
## 18          5.1         3.5          1.4         0.3     setosa     setosa
```

```
## 19          5.7         3.8          1.7         0.3          setosa     setosa
## 20          5.1         3.8          1.5         0.3          setosa     setosa
##       predic[, 2]
## 1   3.896358e-22
## 2   7.217970e-18
## 3   1.463849e-19
## 4   1.268536e-16
## 5   1.637387e-22
## 6   3.883282e-21
## 7   1.113469e-18
## 8   3.877586e-20
## 9   1.902813e-15
## 10  1.111803e-18
## 11  1.185277e-23
## 12  1.621649e-18
## 13  1.459225e-18
## 14  1.117219e-19
## 15  5.487399e-30
## 16  1.261505e-27
## 17  6.754338e-25
## 18  4.223741e-21
## 19  1.774911e-22
## 20  2.593237e-22
```

```
write.csv(datosIr,"tabla con val.predichos iris.csv")
```

```
# tabla de clasificacion (Matriz de confusion)
```

```
addmargins(table(iris$Species,clasific))
```

```
##                        clasific
##              setosa versicolor virginica Sum
## setosa       50     0          0         50
## versicolor   0      48         2         50
## virginica    0      1          49        50
## Sum          50     49         51        150
```

```
prop.table(table(iris$Species,clasific),1)
```

```
##            clasific
##              setosa versicolor virginica
## setosa       1.00   0.00       0.00
## versicolor   0.00   0.96       0.04
## virginica    0.00   0.02       0.98
```

```
predicciones <- predict(object = modelo_lda, newdata = iris[, -5])
table(iris$Species, predicciones$class, dnn = c("Clase real", "Clase predicha"))
```

```
trainig_error <- mean(iris$Species != predicciones$class) * 100
paste("trainig_error =", trainig_error, "%")
```

```
## [1] "trainig_error = 2%"
```

Solo 3 de las 150 predicciones que ha realizado el modelo han sido erróneas. El *trainig error* es muy bajo (2%), lo que apunta a que el modelo es bueno. Sin embargo, para validarlo es necesario un nuevo set de datos con el que calcular el *test error* o recurrir a validación cruzada.

Visualización de las clasificaciones
La función partimat() del paquete klar permite representar los límites de clasificación de un modelo discriminante lineal o cuadrático para cada par de predictores. Cada color representa una región de clasificación acorde al modelo, se muestra el centroide de cada región y el valor real de las observaciones.

```
par(mfcol = c(4, 3))
library(klaR)
partimat(Species ~ Sepal.Width + Sepal.Length + Petal.Length
```

```
    + Petal.Width, data = iris, method = "lda", prec = 200,
    image.colors = c("darkgoldenrod1","snow2",
        "skyblue2"), col.mean = "firebrick")
```