PASOS PARA SU APLICACIÓN

ANÁLISIS DISCRIMINANTE LINEAL

1. Verificar supuestos

1. Los grupos o categorías están bien definidos previamente a iniciar el análisis y se ha verificado que cada uno de las observaciones que servirán para generar la función discriminante han sido correctamente clasificadas en las categorías definidas.
2. Los datos "outliers" deben evitarse ya que el Analisis Discriminate es sensible a ellos.
3. Cuidar que los *grupos de la variable dependiente* (categórica) sean *dos o más y* que estos deben ser *mutuamente excluyentes y exhaustivos*. Pueden crearse grupos "artificiales" cuando se trabaja con variables ordinales o de intervalo.
4. Probar la *normalidad univariada de las variables independientes*. Para lo cual se puede emplear la prueba estadística de normalidad. Si este supuesto es violado se puede utilizar alguna otra técnica multivariada como la regresión logística.

1. Verificar supuestos

5. La *matriz de varianza-covarianza de las variables independientes dentro de cada grupo debe ser igual* (*homogeneidad*). Utilizar la el Test de Igualdad de Matrices de Covarianza para verificarlo. Emplear la prueba *M de Box* que tiene como hipótesis nula que las matrices de covarianza son iguales. En caso de violar el supuesto usar la técnica discriminante cuadrática o agregando más observaciones o excluyendo grupos.

   La presencia de matrices de covariancia no iguales afectan el proceso de clasificación ya que puede clasificar observaciones en los grupos con mayor covarianza. Se puede utilizar la covarianza especifica de cada grupo pero es necesario hacer una validación cruzada del resultado.

6. Tener una *baja multicolinealidad de las variables independientes dentro de cada grupo*. Cuando hay alta multicolinealidad entre dos o mas variables, la función discriminante no asignará correctamente las nuevas observaciones. Utilizar la matriz de correlación para detectar multicolinealidad. En caso de tener valores mayores a 0.8 excluir variables o usar el Análisis de Factores.

2. Estimación del modelo discriminante

Planteamiento:

Sean π_1 y π_2 dos poblaciones normales multivariadas conocidas donde tenemos definidas p variables observables $(X_1, X_2, \ldots, X_p)$ y que las funciones de densidad de ambas poblaciones, f_1 y f_2, es conocidas. Se busca asignar ω a uno de las dos poblaciones.

La regla discriminante es el criterio que permitirá asignar ω conociendo $(x_1, x_2, \ldots, x_p)$ y que a menudo es planteado mediante la *función discriminante* $D(x_1, x_2, \ldots, x_p)$. Entonces la regla de clasificación es:

Si $D(x_1, x_2, \ldots, x_p) > 0$	asignar ω a π_1
En caso contrario	asignar ω a π_2

2. Estimación del modelo discriminante

4 son las formas en que se puede abordar la generación de la regla para dar solución al problema planteado:

A. Regla de verosimilitud
B. Distancia de Mahalanobis
C. Probabilidad *a posteriori*
D. Función discriminante lineal de Fisher

A. Regla de verosimilitud

Para entender el concepto de verosimilitud, suponga que hay tres urnas y en cada urna hay una cantidad diferente de bolas blancas y negras:

Urna	# bolas blancas	# bolas negras
U4	4	96
U50	50	50
U99	99	1

Alguien presenta una muestra de 4 bolas de color blanco y comenta que las tomo de la misma urna. ¿De cuál urna proviene estas bolas? Lo mas probable es que venga de la urna U99 dado existe una mayor cantidad de bolas blancas que negras.

A. Regla de verosimilitud

La regla consiste en asignar ω a donde la verosimilitud de las observaciones x es más grande:

Si $f_1(x; \mu_1, \sum_1) > f_2(x; \mu_2, \sum_2)$ asignar ω a π_1
En caso contrario asignar ω a π_2

En donde $f_i(x; \mu_i, \sum_i)$ representa la función de verosimilitud para la i-ésima población.

La función de verosimilitud (probabilidad) es la función de densidad de verosimilitud normal multivariada.

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\right\}$$

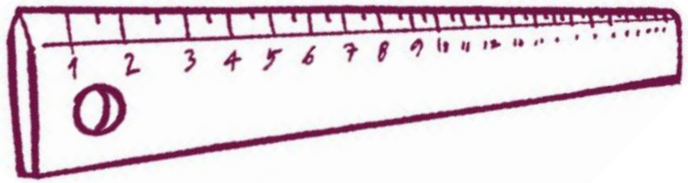
B. Distancia de Mahalanobis

Cuando $\sum_1 = \sum_2$ la regla de verosimilitud es equivalente a decir:

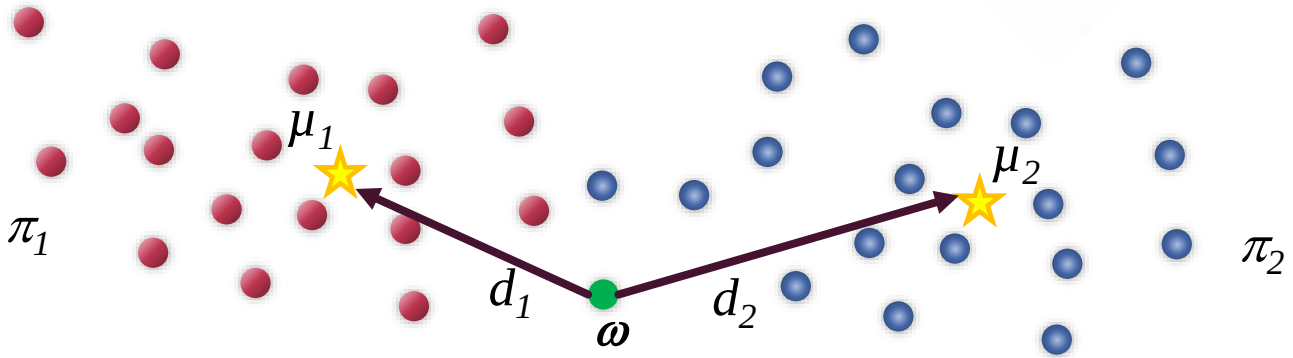
Si $d_1 < d_2$	asignar ω a π_1
En caso contrario	asignar ω a π_2

En donde

$$d_i = [\omega - \mu_i]'\Sigma^{-1}[\omega - \mu_i]$$



La cantidad d_i, llamada cuadrado de la distancia de Mahalanobis, es una medida de lo lejos que está ω de μ_i.



Ejemplo

Mytilicola intestinalis es un parásito del mejillón, que en estado larval presenta diferentes estadios de crecimiento. El primer estadio (*Nauplis*) y el segundo estadio (*Metanauplius*) son difíciles de distinguir. Sobre una muestra de $n_1=76$ y $n_2=91$ copépodos que se pudieron identificar al microscopio como del primero y segundo estadio respectivamente, se midieron las variables l = *longitud*, a = *anchura*, y se obtuvieron las siguientes medias y matrices de covarianzas:

$$\bar{X}_1 = \begin{bmatrix} 219.5 \\ 138.1 \end{bmatrix} \qquad \bar{X}_2 = \begin{bmatrix} 241.6 \\ 147.8 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 409.9 & -1.316 \\ -1.316 & 306.2 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 201.9 & 57.97 \\ 57.97 & 152.8 \end{bmatrix}$$

La matriz de covarianza común es:

$$S = \frac{n_1 S_1 + n_2 S_2}{n_1 + n_2} = \frac{76 \begin{bmatrix} 409.9 & -1.316 \\ -1.316 & 306.2 \end{bmatrix} + 91 \begin{bmatrix} 201.9 & 57.97 \\ 57.97 & 152.8 \end{bmatrix}}{76 + 91} = \begin{bmatrix} 296.6 & 30.99 \\ 30.99 & 222.6 \end{bmatrix}$$

Ejemplo

$$d_1 = \omega'\Sigma^{-1}\omega + \mu_1'\Sigma^{-1}\mu_1 - 2\omega'\Sigma^{-1}\mu_1$$

$$d_1 = \begin{bmatrix} l & a \end{bmatrix}\begin{bmatrix} 0.0034 & -0.0005 \\ -0.0005 & 0.0046 \end{bmatrix}\begin{bmatrix} l \\ a \end{bmatrix} + \begin{bmatrix} 219.5 & 138.1 \end{bmatrix}\begin{bmatrix} 0.0034 & -0.0005 \\ -0.0005 & 0.0046 \end{bmatrix}\begin{bmatrix} 219.5 \\ 138.1 \end{bmatrix} - 2\begin{bmatrix} l & a \end{bmatrix}\begin{bmatrix} 0.0034 & -0.0005 \\ -0.0005 & 0.0046 \end{bmatrix}\begin{bmatrix} 219.5 \\ 138.1 \end{bmatrix}$$

$$d_1 = (0.0034l^2 - 0.0010la + 0.0046a^2) + 222.90 - (1.370l + 1.050a)$$

$$d_1 = 0.0034l^2 - 1.370l - 0.0010la - 1.050a + 0.0046a^2 + 222.90$$

$$d_2 = 0.0034l^2 - 1.512l - 0.0010la - 1.117a + 0.0046a^2 + 265.27$$

Ejemplo

Se desea clasificar el siguiente registro: $l = 223.1 \qquad a = 141.2$

$$d_1 = 0.0034l^2 - 1.370l - 0.0010la - 1.050a + 0.0046a^2 + 222.90$$
$$d_1 = -1.5662$$

$$d_2 = 0.0034l^2 - 1.512l - 0.0010la - 1.117a + 0.0046a^2 + 265.27$$
$$d_2 = -0.3368$$

Regla: Si $d_1 < d_2$ asignar ω a π_1

Conclusión: *el registro pertenece al estadio I*

C. Probabilidad *a posteriori*

Si $\sum_1 = \sum_2$ y conocemos las probabilidades *a priori* Ω_1 y Ω_2, con $\Omega_1+\Omega_2 = 1$, de que el elemento venga de cada una de las dos poblaciones.

Una vez observado ω podemos calcular las probabilidades *a posteriori* de que el elemento haya sido generado por cada una de las dos poblaciones, $P(\pi_i|\omega)$, para $i = 1, 2$, por medio del *Teorema de Bayes*.

Por lo que tenemos que:

$$P(\pi_1|\omega) = \frac{f_1(\omega)\Omega_1}{f_1(\omega)\Omega_1 + f_2(\omega)\Omega_2} \longrightarrow P(\pi_1|\omega) = \frac{e^{-\frac{1}{2}d_1}}{e^{-\frac{1}{2}d_1} + e^{-\frac{1}{2}d_2}}$$

$$P(\pi_2|\omega) = \frac{f_2(\omega)\Omega_2}{f_1(\omega)\Omega_1 + f_2(\omega)\Omega_2} \longrightarrow P(\pi_2|\omega) = \frac{e^{-\frac{1}{2}d_2}}{e^{-\frac{1}{2}d_1} + e^{-\frac{1}{2}d_2}}$$

C. Probabilidad *a posteriori*

La regla discriminante será:

Si $P(\pi_1\|\omega) > P(\pi_2\|\omega)$	asignar ω a π_1
En caso contrario	asignar ω a π_2

D. Función discriminante lineal de Fisher

Propone transformar las observaciones multivariadas provenientes de las poblaciones π_1 y π_2 a observaciones univariadas y de tal forma que estas queden separadas lo más posible mediante una combinación lineal.

$$y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

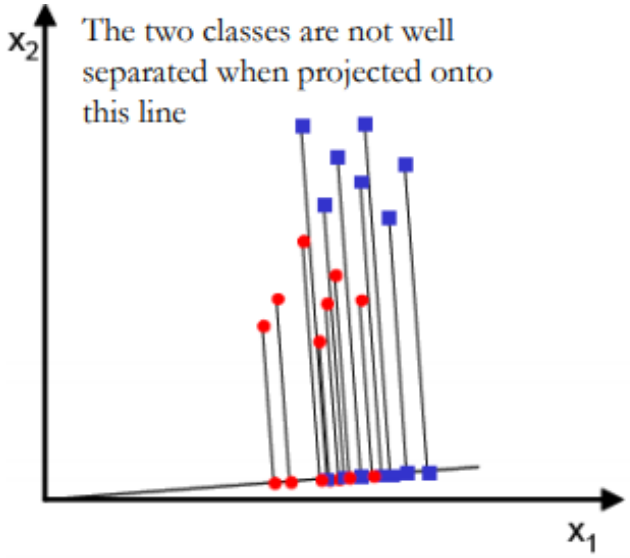
Siendo los valores $y_{11}, y_{12}, \ldots, y_{1n_1}$ las observaciones para la primera población y los valores $y_{21}, y_{22}, \ldots, y_{2n_2}$ para la segunda población. Además de tener matrices iguales varianza-covarianza ($\sum_1 = \sum_2$). La separación puede ser:

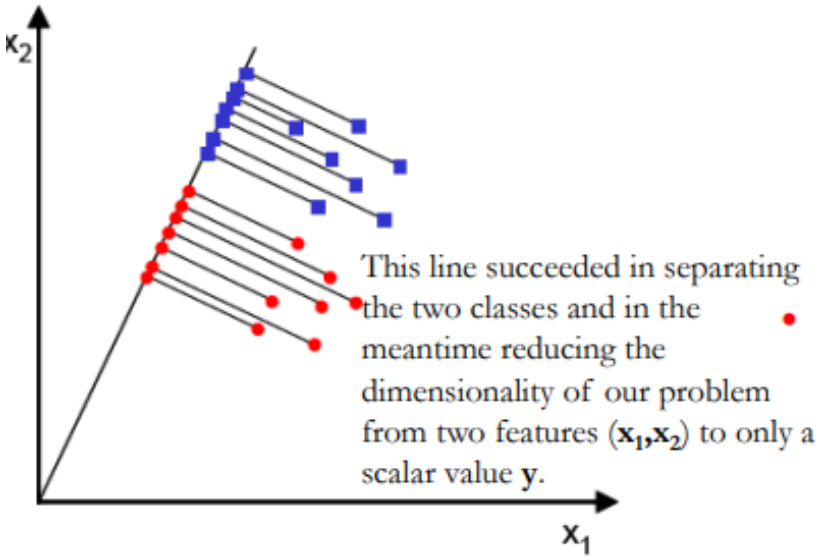
Si $b'X - k > 0$	asignar ω a π_1
En caso contrario	asignar ω a π_2

Siendo:

$$b = \Sigma^{-1}[\mu_1 - \mu_2] \qquad k = \frac{1}{2}[\mu_1 + \mu_2]'\Sigma^{-1}[\mu_1 - \mu_2]$$

D. Función discriminante lineal de Fisher





El método busca la combinación lineal (función discriminante) que logre una mejor separación entre los grupos

Ejemplo: *Mytilicola intestinalis*

$$b = \Sigma^{-1}[\mu_1 - \mu_2]$$

$$b = \begin{bmatrix} 0.0034 & -0.0005 \\ -0.0005 & 0.0046 \end{bmatrix} \begin{bmatrix} 219.5 - 241.6 \\ 138.1 - 147.8 \end{bmatrix}$$

$$b = \begin{bmatrix} -0.071 \\ -0.034 \end{bmatrix}$$

$$k = \frac{1}{2}[\mu_1 + \mu_2]'\Sigma^{-1}[\mu_1 - \mu_2]$$

$$k = \frac{1}{2}[219.5 + 241.6 \quad 138.1 + 147.8] \begin{bmatrix} 0.0034 & -0.0005 \\ -0.0005 & 0.0046 \end{bmatrix} \begin{bmatrix} 219.5 - 241.6 \\ 138.1 - 147.8 \end{bmatrix}$$

$$k = -21.183$$

$$b'X - k$$

$$[-0.071 \quad -0.034] \begin{bmatrix} l \\ a \end{bmatrix} - (-21.183) = \boxed{-0.071l - 0.034a + 21.183}$$

Ejemplo: *Mytilicola intestinalis*

Se desea clasificar el siguiente registro: $l = 223.1 \qquad a = 141.2$

Regla: Si $b'X - k > 0$ asignar ω a π_1

$$-0.071l - 0.034a + 21.183 > 0$$
$$-0.071(223.1) - 0.034(141.2) + 21.183 > 0$$

$$0.5421 > 0$$

Conclusión: *el registro pertenece al estadio I*

Ejemplo: *Mytilicola intestinalis*

Además, retomando las ecuaciones de la distancia de Mahalanobis

$$0 = d_2 - d_1$$

$$0 = (0.0034l^2 - 1.512l - 0.0010la - 1.117a + 0.0046a^2 + 265.27\ ) \\ -(0.0034l^2 - 1.370l - 0.0010la - 1.050a + 0.0046a^2 + 222.90)$$

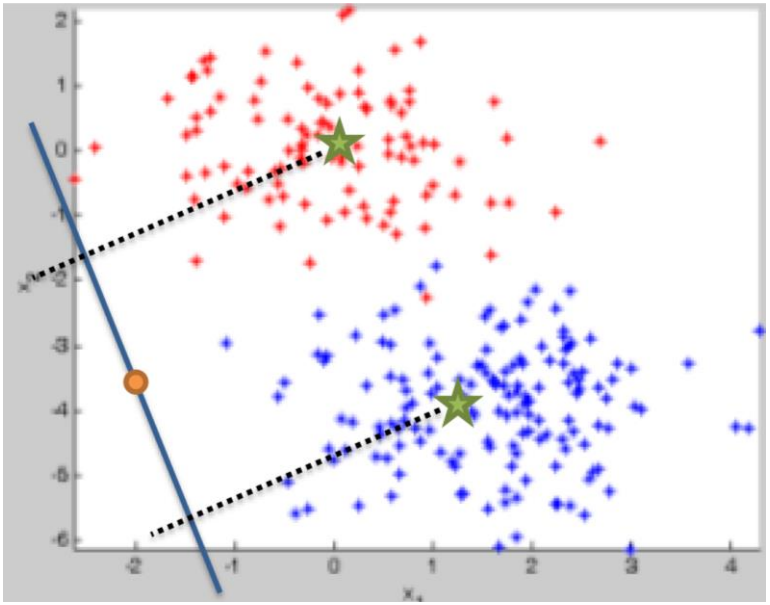
$$0 = -0.1420l - 0.0674a + 42.3666$$

$$0 < \frac{1}{2}[-0.1420l - 0.0674a + 42.3666]$$

$$0 < -0.0710l - 0.0332a + 21.183$$

$$-0.0710l - 0.0332a + 21.183 > 0$$

Función discriminante de Fisher



Consideración de las consecuencias

Considere que una máquina automática clasifica equivocadamente un billete de 20 pesos como uno de 50, y devuelve el cambio equivocado, el costo es de 30 pesos.

La consecuencia puede resultar no tan grave, pero si un paciente tuviera que entrar a cirugía y al momento de ponerle la anestesia tiene una complicación a pesar de haber sido clasificado como apto a recibirla. ¿Cómo consideraría el resultado?

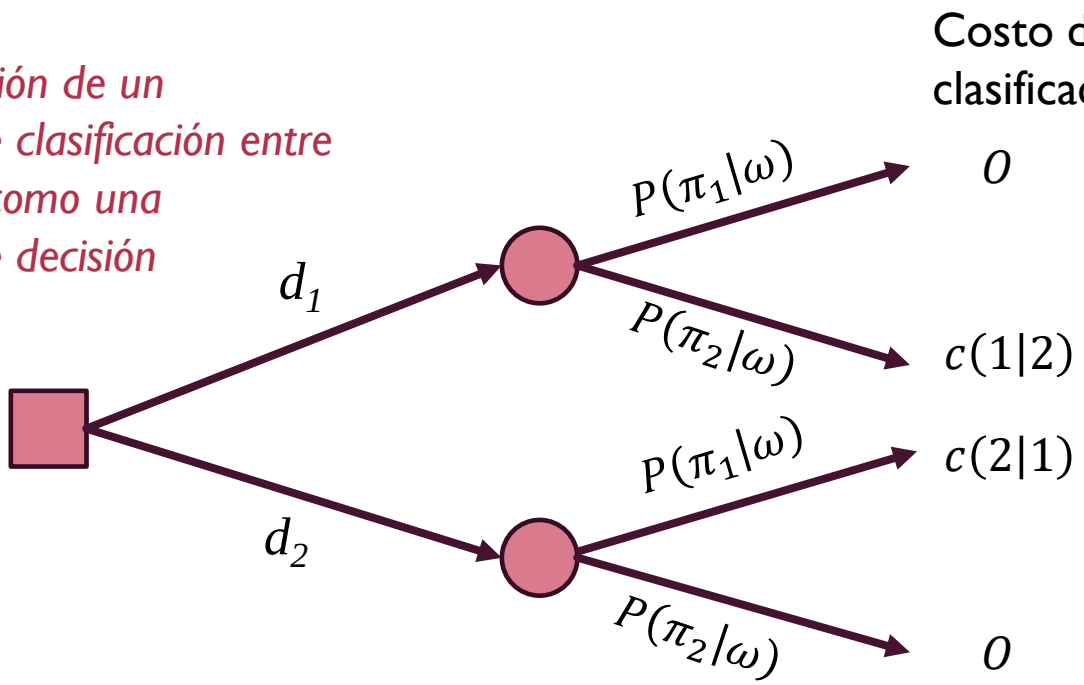
Si las consecuencias pueden cuantificarse, se pueden incluir para el calculo de la función discriminante formulándolo como un problema bayesiano de decisión. Supongamos que:

1. $c(i|j)$ es el costo conocido de clasificar en π_i una unidad que pertenece a π_j.
2. El decisor quiere maximizar su función de utilidad o minimizar el costo esperado.

Consideración de las consecuencias

Con estas dos hipótesis la mejor decisión es la que minimiza los costos esperados.

Representación de un problema de clasificación entre dos grupos como una problema de decisión



La nueva regla de clasificación quedaría:

Si $c(1\|2)P(\pi_2\|\omega) < c(2\|1)P(\pi_1\|\omega)$	asiganar ω a π_1
En caso contrario	asiganar ω a π_2

Consideración de las consecuencias

Si se cumple que $\sum_1 = \sum_2$ entonces clasificaremos ω a π_1 si:

a) Su probabilidad *a priori* es más alta
b) La verosimilitud de que ω provenga de π_1 es más alta;
c) El costo de equivocarnos al clasificarlo en π_1 es más bajo.

La regla de clasificación quedaría:

Si $b'X - k > ln\left[\left(\frac{c(1\|2)}{c(2\|1)}\right)\left(\frac{\Omega_2}{\Omega_1}\right)\right]$	asignar ω a π_1
En caso contrario	asignar ω a π_2

Ejemplo: *Mytilicola intestinalis*

Suponga que se conoce la probabilidad a priori de la larva se encuentre en el estadío 1 que es de 0.25. Asumiendo que el costo de una clasificación errónea es de 0.50 para π_1 y de 1.5 para π_2. Donde debería clasificarse $l = 223.1; a = 141.2$

Ejemplo: *Mytilicola intestinalis*

Regla: Si $b'X - k > ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{\Omega_2}{\Omega_1}\right)\right]$ asignar ω a π_1

$$-0.071l - 0.034a + 21.183 > ln\left[\left(\frac{1.50}{0.50}\right)\left(\frac{0.75}{0.25}\right)\right]$$

$$-0.071(223.1) - 0.034(141.2) + 21.183 > ln[9]$$

$0.5421 > 2.20$

Conclusión: *el registro pertenece al estadio 2*

Selección de variables

- ¿Son necesarias todas las variables para un análisis eficaz?
- ¿Cuáles son las mejores variables para discriminar?

Métodos para seleccionar variables:

1) Selección hacia adelante (*forward stepwise*)
2) Eliminación hacia atrás (*backward stepwise*)
3) Selección por paso (combinación de los anteriores)

La capacidad discriminante de la variable independiente puede ser descrita por los valores parciales de F. Valores de la F grandes indican una capacidad discriminante mayor.

3. Valorar la exactitud de la predicción

Dado que la variable dependiente es no métrica, para realizar la validación debe valorarse cada observación como si fuera correctamente clasificada.

El procedimiento más frecuentemente utilizado es dividir en dos grupos aleatoriamente en la muestra de análisis y en una ampliación de la muestra. El primer grupo sirve para generar la función discriminante mientras que el segundo validará el número de casos bien clasificados.

Otro procedimiento es realizar una validación cruzada basados en el principio de "dejar uno afuera" donde la función discriminante es ajustada con muestras tomadas repetidamente de la original. Elimina una observación de las muestra.