CEDEX - Curso de formación estadística

# Práctica 1

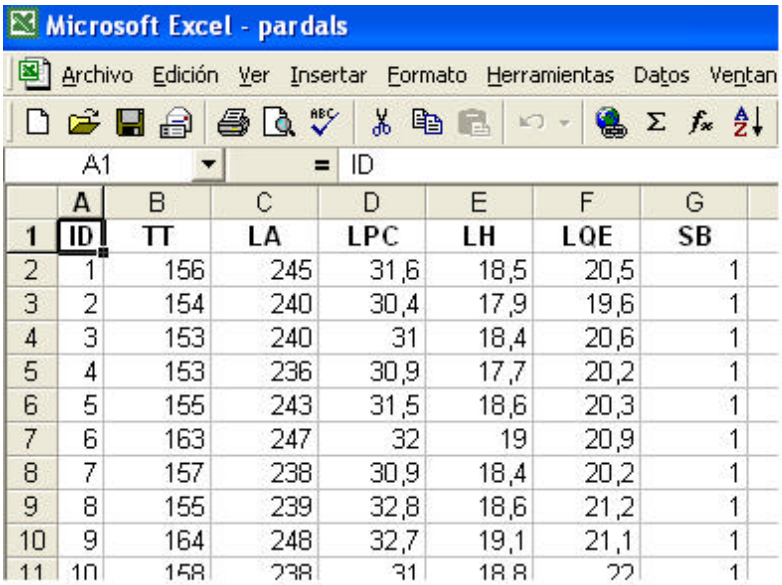
## 1. Introducción

En esta práctica trabajaremos con el fichero de datos `pardals.xls` que corresponde a un estudio de variables biométricas realizado por Cuadras y colaboradores. El fichero contiene medidas de 49 pájaros recogidos casi moribundos después de un temporal. Las variables son:

| Nombre | Descripción |
|---|---|
| tt | tamaño total |
| la | longitud del ala |
| lpc | longitud del pico y cabeza |
| lh | longitud del húmero |
| lqe | longitud de la quilla del esternón |
| sb | (= 1) si sobrevive (= 0) si no sobrevive |

### 1.1. ¿Cómo introducir datos en SPSS/PC?

A continuación se ilustra como introducir los datos utilizando el fichero de datos en formato Excel: `pardals.xls`. La estructura de este fichero es la siguiente:




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | TT | LA | LPC | LH | LQE | SB |
| 2 | 1 | 156 | 245 | 31,6 | 18,5 | 20,5 | 1 |
| 3 | 2 | 154 | 240 | 30,4 | 17,9 | 19,6 | 1 |
| 4 | 3 | 153 | 240 | 31 | 18,4 | 20,6 | 1 |
| 5 | 4 | 153 | 236 | 30,9 | 17,7 | 20,2 | 1 |
| 6 | 5 | 155 | 243 | 31,5 | 18,6 | 20,3 | 1 |
| 7 | 6 | 163 | 247 | 32 | 19 | 20,9 | 1 |
| 8 | 7 | 157 | 238 | 30,9 | 18,4 | 20,2 | 1 |
| 9 | 8 | 155 | 239 | 32,8 | 18,6 | 21,2 | 1 |
| 10 | 9 | 164 | 248 | 32,7 | 19,1 | 21,1 | 1 |
| 11 | 10 | 158 | 238 | 31 | 18,8 | 22 | 1 |

Notemos que en este fichero aparecen los nombres de las variables. Para importar los datos se procede como sigue:

1. `Archivo` ↪ `Abrir` ↪ `Datos`.
2. En la ventana `Abrir Archivo`, en `Tipos de archivos:` elegir la extensión `Excel (*.xls)`, y seleccionar el archivo `pardals.xls` ↪ `Abrir`.
3. En la ventana `Apertura de fuente de datos de Excel` marcar la opción ☑ `Leer nombre de variables de la primera fila de datos` ↪ `Aceptar`.
4. Por último, `Archivo` ↪ `Guardar`.

## 2. Análisis exploratorio

El análisis exploratorio es la primera fase de cualquier estudio de datos y puesto que nos enfrentamos por primera vez al conjunto de datos `pardals.xls` "repetiremos" la práctica 1 con estos datos.

### 2.1. Estadísticos descriptivos univariantes

Obtenemos los estadísticos descriptivos (media, mediana, moda, varianza, etc.) por categoría (sobrevivencia):

1. `Analizar` ↪ `Estadísticos descriptivos` ↪ `Explorar ...`.
2. Pasar las variables cuantitativas del recuadro izquierdo al recuadro derecho `Dependientes:`.
3. Pasar la variable `sb` del recuadro izquierdo al recuadro derecho `Factores:`.
4. Por último, ↪ `Aceptar`.

Los resultados de la secuencia anterior (presentamos sólo la primera variable) son:

| | SB | | | Estadístico | Error típ. |
|---|---|---|---|---|---|
| TT | 0 | Media | | 158,43 | ,734 |
| | | Intervalo de confianza para la media al 95% | Límite inferior | 156,92 | |
| | | | Límite superior | 159,93 | |
| | | Media recortada al 5% | | 158,44 | |
| | | Mediana | | 159,00 | |
| | | Varianza | | 15,069 | |
| | | Desv. típ. | | 3,882 | |
| | | Mínimo | | 152 | |
| | | Máximo | | 165 | |
| | | Rango | | 13 | |
| | | Amplitud intercuartil | | 7,00 | |
| | | Asimetría | | -,125 | ,441 |
| | | Curtosis | | -1,227 | ,858 |
| | 1 | Media | | 157,38 | ,725 |
| | | Intervalo de confianza para la media al 95% | Límite inferior | 155,87 | |
| | | | Límite superior | 158,89 | |
| | | Media recortada al 5% | | 157,26 | |
| | | Mediana | | 157,00 | |
| | | Varianza | | 11,048 | |
| | | Desv. típ. | | 3,324 | |
| | | Mínimo | | 153 | |
| | | Máximo | | 164 | |
| | | Rango | | 11 | |
| | | Amplitud intercuartil | | 4,50 | |
| | | Asimetría | | ,569 | ,501 |
| | | Curtosis | | -,477 | ,972 |

## 2.2. Matriz de correlaciones

Obtenemos la matriz de correlaciones:

1. `Analizar` ↪ `Correlaciones` ↪ `Bivariadas ...`.
2. Pasar todas las variables cuantitativas del recuadro izquierdo al recuadro derecho `Variables`.
3. Por último, ↪ `Aceptar`.

**Correlaciones**

| | | TT | LA | LPC | LH | LQE |
|---|---|---|---|---|---|---|
| TT | Correlación de Pearson | 1 | ,735** | ,662** | ,645** | ,605** |
| | Sig. (bilateral) | , | ,000 | ,000 | ,000 | ,000 |
| | N | 49 | 49 | 49 | 49 | 49 |
| LA | Correlación de Pearson | ,735** | 1 | ,674** | ,769** | ,529** |
| | Sig. (bilateral) | ,000 | , | ,000 | ,000 | ,000 |
| | N | 49 | 49 | 49 | 49 | 49 |
| LPC | Correlación de Pearson | ,662** | ,674** | 1 | ,763** | ,526** |
| | Sig. (bilateral) | ,000 | ,000 | , | ,000 | ,000 |
| | N | 49 | 49 | 49 | 49 | 49 |
| LH | Correlación de Pearson | ,645** | ,769** | ,763** | 1 | ,607** |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | , | ,000 |
| | N | 49 | 49 | 49 | 49 | 49 |
| LQE | Correlación de Pearson | ,605** | ,529** | ,526** | ,607** | 1 |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | ,000 | , |
| | N | 49 | 49 | 49 | 49 | 49 |

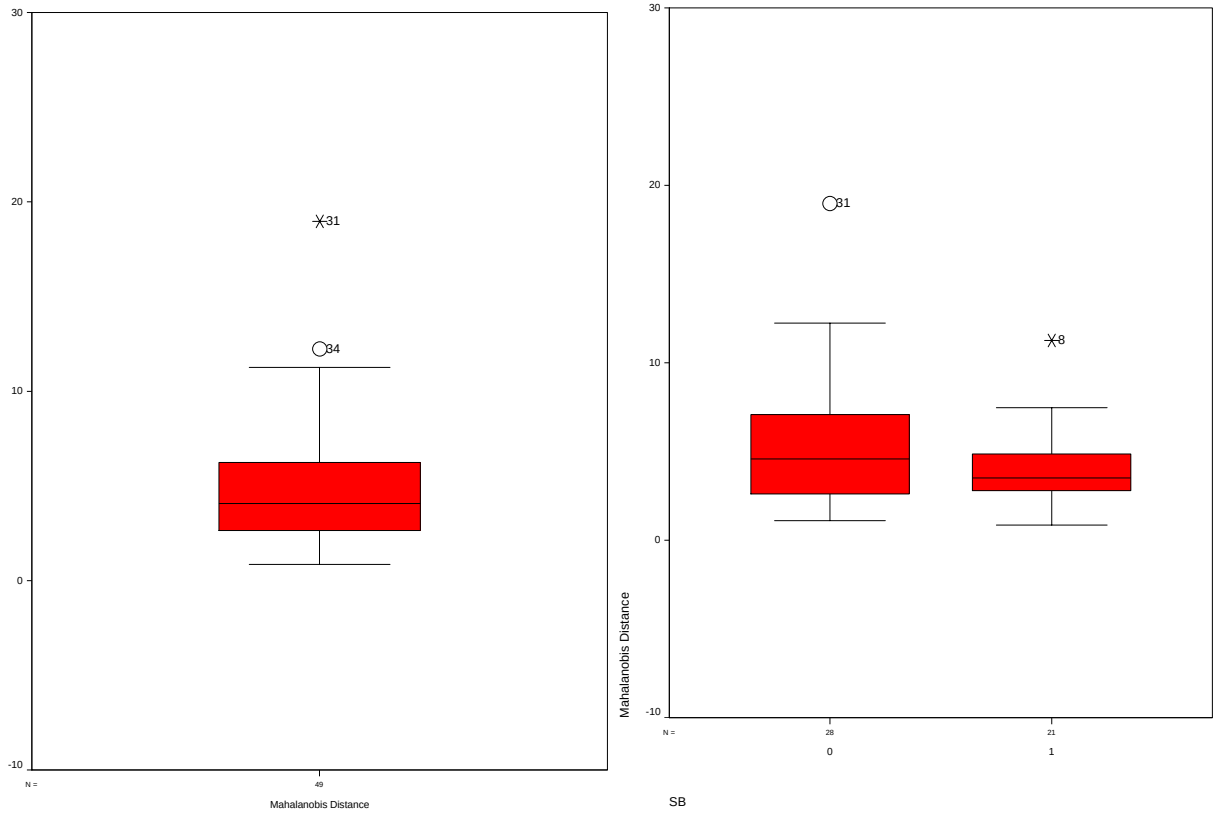
**. La correlación es significativa al nivel 0,01 (bilateral).

## 2.3. Distancia de Mahalanobis

Obtenemos la distancia de Mahalanobis:

1. `Analizar` ↪ `Regresión` ↪ `Lineal ...`.
2. Pasar la variable `sb` al recuadro `Dependiente:`.
3. Pasar todas las variables cuantitativas del recuadro izquierdo al recuadro `Independientes:`.
4. Seleccionar `Guardar ...`, marcar ☑ `Mahalanobis` ↪ `Continuar`.
5. Por último, ↪ `Aceptar`.

Los gráficos siguientes presentan los diagramas de caja de la variable `mah_1` en la muestra completa y por categorías:




Ejercicio 1. En base a los resultados obtenidos en esta sección responda la Pregunta 1 de esta práctica.

## 3. Análisis de componentes principales - I

Para obtener las componentes principales utilizando la matriz de covarianzas seguimos la siguiente secuencia de pasos:

1. `Analizar` ↪ `Reducción de datos` ↪ `Análisis factorial ...`.
2. Pasar las variables cuantitativas al recuadro `Variables:`.
3. Seleccionamos `Extracción ...`:
   *a*) En `Método:` seleccionamos la opción `Componentes principales`.
   *b*) Marcamos la opción: ⊙ `Matriz de covarianza`.
   *c*) Marcamos la opción: ☑ `Gráfico de sedimentación`.
   *d*) En `Autovalores mayores que:` escribimos `1` ↪ `Continuar`.
4. Seleccionamos `Puntuaciones ...`:
   *a*) Marcamos la opción: ☑ `Guardar como variables`.
   *b*) Marcamos la opción: ⊙ `Regresión` ↪ `Continuar`.
5. Por último, ↪ `Aceptar`.

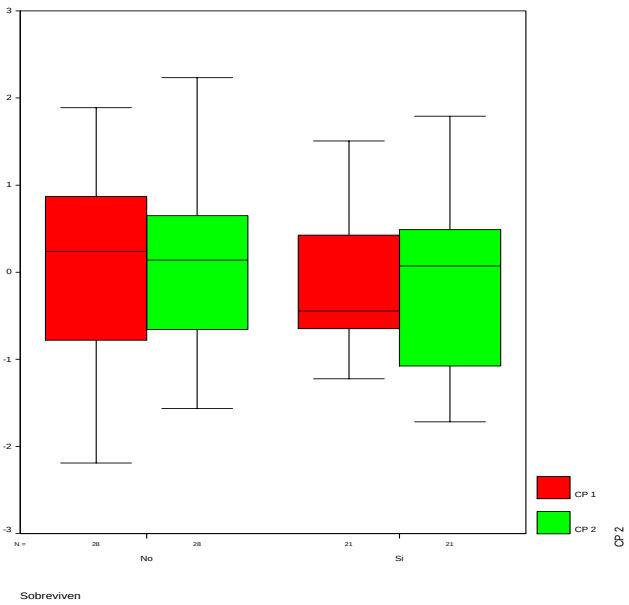
Como resultado tenemos un análisis de componentes principales con una componente principal, `fac1_1`, que explica el 86.23 % de la variabilidad total y que puede interpretarse como una componente de tamaño.

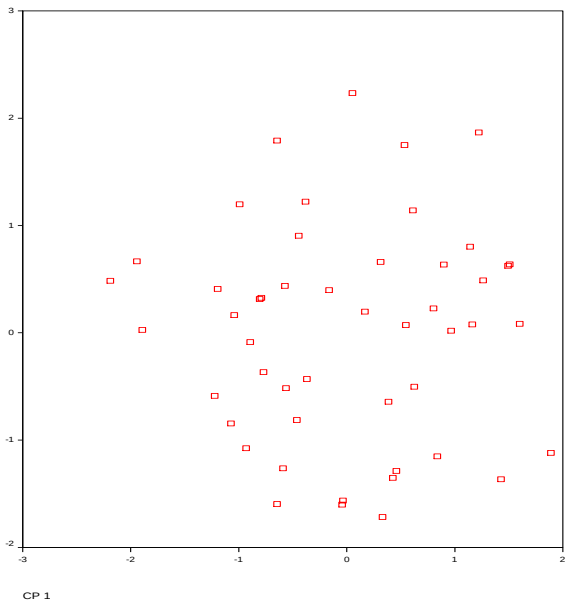
| | | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|---|
| | Componente | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| Bruta | 1 | 35,326 | 86,225 | 86,225 | 35,326 | 86,225 | 86,225 |
| | 2 | 4,622 | 11,283 | 97,507 | | | |
| | 3 | ,631 | 1,540 | 99,047 | | | |
| | 4 | ,313 | ,763 | 99,811 | | | |
| | 5 | 7,752E-02 | ,189 | 100,000 | | | |

| | Bruta | Reescalada |
|---|---|---|
| | Componente | Componente |
| | 1 | 1 |
| TT | 3,189 | ,873 |
| LA | 4,927 | ,972 |
| LPC | ,574 | ,722 |
| LH | ,442 | ,783 |
| LQE | ,596 | ,601 |

EJERCICIO 2. Obtenga un análisis de componentes principales que explique más del 90 % de la variabilidad total y responda la PREGUNTA 2 de esta práctica.

EJERCICIO 3. Obtenga los diagramas de cajas por sobrevivencia y la matriz de diagramas de dispersión en la muestra completa de las componentes principales calculadas en el ejercicio anterior y responda la PREGUNTA 3 de esta práctica.




## 4. Análisis de componentes principales - II

Para obtener las componentes principales utilizando la matriz de correlaciones seguimos la siguiente secuencia de pasos:

1. `Analizar` ↪ `Reducción de datos` ↪ `Análisis factorial ...`.
2. Pasar las variables cuantitativas al recuadro `Variables:`.
3. Seleccionamos `Extracción ...`:
   - *a*) En `Método:` seleccionamos la opción `Componentes principales`.
   - *b*) Marcamos la opción: ⊙ `Matriz de correlación`.
   - *c*) Marcamos la opción: ☑ `Gráfico de sedimentación`.
   - *d*) En `Numero de factores:` escribimos `3` ↪ `Continuar`.
4. Seleccionamos `Puntuaciones ...`:
   - *a*) Marcamos la opción: ☑ `Guardar como variables`.
   - *b*) Marcamos la opción: ⊙ `Regresión` ↪ `Continuar`.
5. Por último, ↪ `Aceptar`.

Como resultado tenemos un análisis de componentes principales con tres componentes principales, `fac1_3`, `fac2_3` y `fac3_3`, que explican el 90.68 % de la variabilidad total.

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 3,616 | 72,320 | 72,320 | 3,616 | 72,320 | 72,320 |
| 2 | ,532 | 10,630 | 82,950 | ,532 | 10,630 | 82,950 |
| 3 | ,386 | 7,728 | 90,678 | ,386 | 7,728 | 90,678 |
| 4 | ,302 | 6,031 | 96,709 | | | |
| 5 | ,165 | 3,291 | 100,000 | | | |

| | Componente | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| TT | ,859 | 3,698E-02 | -,429 |
| LA | ,878 | -,218 | -,212 |
| LPC | ,857 | -,237 | ,283 |
| LH | ,895 | -,135 | ,255 |
| LQE | ,756 | ,639 | ,111 |

EJERCICIO 4. Obtenga los diagramas de cajas por sobrevivencia y la matriz de diagramas de dispersión en la muestra completa de las tres componentes principales y responda la PREGUNTA 4 de esta práctica.

## 5. Análisis factorial

Comenzaremos por obtener un modelo factorial con un único factor común:

1. `Analizar` ↪ `Reducción de datos` ↪ `Análisis factorial ...`.
2. Pasar las variables cuantitativas al recuadro `Variables:`.
3. Seleccionamos `Descriptivos ...`:
   - *a*) Marcamos la opción: ☑ `Coeficientes`.
   - *b*) Marcamos la opción: ☑ `Reproducida` ↪ `Continuar`.
4. Seleccionamos `Extracción ...`:
   - *a*) En `Método:` seleccionamos la opción `Máxima verosimilitud`.
   - *b*) Marcamos la opción: ☑ `Gráfico de sedimentación`.
   - *c*) En `Numero de factores:` escribimos `1` ↪ `Continuar`.
5. Seleccionamos `Puntuaciones ...`:
   - *a*) Marcamos la opción: ☑ `Guardar como variables`.
   - *b*) Marcamos la opción: ⊙ `Regresión` ↪ `Continuar`.
6. Por último, ↪ `Aceptar`.

| Factor | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 3,616 | 72,320 | 72,320 | 3,286 | 65,728 | 65,728 |
| 2 | ,532 | 10,630 | 82,950 | | | |
| 3 | ,386 | 7,728 | 90,678 | | | |
| 4 | ,302 | 6,031 | 96,709 | | | |
| 5 | ,165 | 3,291 | 100,000 | | | |

Método de extracción: Máxima verosimilitud.

**Matriz factorial[a]**

| | Factor |
|---|---|
| | 1 |
| TT | ,800 |
| LA | ,859 |
| LPC | ,825 |
| LH | ,884 |
| LQE | ,669 |

Método de extracción: Máxima verosimilitud.

a. 1 factores extraídos. Requeridas 4 iteraciones.

**Prueba de la bondad de ajuste**

| Chi-cuadrado | gl | Sig. |
|---|---|---|
| 8,143 | 5 | ,149 |

Vimos en la sección anterior que la matriz de correlaciones era:

**Matriz de correlaciones**

| | | TT | LA | LPC | LH | LQE |
|---|---|---|---|---|---|---|
| Correlación | TT | 1,000 | ,735 | ,662 | ,645 | ,605 |
| | LA | ,735 | 1,000 | ,674 | ,769 | ,529 |
| | LPC | ,662 | ,674 | 1,000 | ,763 | ,526 |
| | LH | ,645 | ,769 | ,763 | 1,000 | ,607 |
| | LQE | ,605 | ,529 | ,526 | ,607 | 1,000 |

y sabemos que el modelo factorial cumple que $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$. Por tanto, si el modelo factorial es satisfactorio podemos reconstruir la matriz de correlaciones a partir de las matrices $\mathbf{\Lambda}$ y $\mathbf{\Psi}$:

**Correlaciones reproducidas**

| | | TT | LA | LPC | LH | LQE |
|---|---|---|---|---|---|---|
| Correlación reproducida | TT | ,639[b] | ,687 | ,659 | ,707 | ,535 |
| | LA | ,687 | ,738[b] | ,708 | ,759 | ,575 |
| | LPC | ,659 | ,708 | ,680[b] | ,729 | ,552 |
| | LH | ,707 | ,759 | ,729 | ,782[b] | ,592 |
| | LQE | ,535 | ,575 | ,552 | ,592 | ,448[b] |
| Residual[a] | TT | | 4,825E-02 | 2,482E-03 | -6,158E-02 | 7,001E-02 |
| | LA | 4,825E-02 | | -3,448E-02 | 9,235E-03 | -4,578E-02 |
| | LPC | 2,482E-03 | -3,448E-02 | | 3,420E-02 | -2,560E-02 |
| | LH | -6,158E-02 | 9,235E-03 | 3,420E-02 | | 1,499E-02 |
| | LQE | 7,001E-02 | -4,578E-02 | -2,560E-02 | 1,499E-02 | |

Método de extracción: Máxima verosimilitud.

a. Los residuos se calculan entre las correlaciones observadas y reproducidas. Hay 2 (20,0%) residuales no redundantes con valores absolutos mayores que 0,05.

b. Comunalidades reproducidas

EJERCICIO 5. En base los resultados obtenidos responda la PREGUNTA 5 de esta práctica.

# Técnicas de análisis multivariante - I

Nombre y Apellidos:

**1**. ¿Cuál de las siguientes afirmaciones es falsa?

- ◯ En la muestra completa, el máximo de la variable tamaño total, `tt`, es igual a 165 y se alcanza en un pájaro que no sobrevive.
- ◯ La observación 31, que corresponde a un pájaro que no sobrevive, es un posible atípico en la muestra.
- ◯ La observación 8, que corresponde a un pájaro que sobrevive, es un posible atípico en la muestra.

**2**. ¿Cuál de las siguientes afirmaciones es falsa?

- ◯ La segunda componente contrapone las medidas de las extremidades al resto de las medidas.
- ◯ La componente principal, `fac1_2`, es incorrelada con la segunda componente principal, `fac2_2`.
- ◯ La primera componente principal del segundo análisis es igual, casualmente, a la componente principal del primer análisis.

**3**. ¿Cuál de las siguientes afirmaciones es falsa?

- ◯ El máximo y el mínimo de la primera componente principal, `fac1_2`, se obtienen en especímenes que no sobreviven.
- ◯ Los menores valores de la primera componente principal, `fac1_2`, se alcanzan en las observaciones 25, 30 y 37.
- ◯ La segunda componente principal, `fac2_2`, crece linealmente con el aumento de la primera componente principal, `fac1_2`.

**4**. ¿Cuál de las siguientes afirmaciones es falsa?

- ◯ La correlación entre la primera componente principal, `fac1_3`, y la primera variable estandarizada, `ztt`, es igual a `0.859`.
- ◯ La variabilidad explicada por la primera componente principal basada en la matriz de covarianzas puede coincidir con la variabilidad explicada por la primera componente principal normada.
- ◯ La variabilidad explicada por la primera componente principal basada en la matriz de covarianzas nunca coincide con la variabilidad explicada por la primera componente principal normada.

**5**. ¿Cuál de las siguientes afirmaciones es falsa?

- ◯ La estimación de la matriz de carga utilizando el método de las componentes principales es, en general, diferente a la estimación utilizando el método de máxima verosimilitud.
- ◯ La proporción de la variabilidad total explicada por el primer factor es igual a 72,32 %.
- ◯ En el contraste de bondad de ajuste se rechaza la hipótesis nula, $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$.

**6**. En un centro de acogida de aves tienen 100 ejemplares adultos de la misma especie que la considerada en este estudio y desean liberar 80 ejemplares y conservar 20 ejemplares para que se reproduzcan en cautividad. Los ejemplares que se liberarán, deberían ser los que tengan mayor probabilidad de sobrevivir. ¿Teniendo en cuenta los resultados obtenidos en la práctica, que recomendarías acerca de las características biométricas de los ejemplares a liberar?