ESTADÍSTICA GRADO EN CIENCIAS DEL MAR, CURSO 2011-12

ANÁLISIS DE DATOS - TAREA Nº 1

Incluir a continuación el DNI del alumno que realiza esta práctica:

ALUMNO

Calificación: (GLOBAL)

8.67

Figura~1: Crías de tortuga dirigiéndose al mar tras emerger del nido

Entre los años 1999 y 2004 se llevaron a cabo diversas campañas para el estudio del anidamiento y éxito reproductivo de la tortuga boba (*Caretta caretta*) en varias playas de la isla de Boavista en el archipiélago de Cabo Verde. Como resultado de esta campaña se recogieron datos que se han guardado en un archivo .csv (*comma separated values*). Los datos se han organizado de tal forma que cada fila del archivo corresponde a una tortuga de esta especie, con información adicional sobre el nido construido por ese ejemplar si la tortuga es hembra. Para cada animal, las variables que se han medido son:

Anno: Año de la toma de datos.

Periodo: Quincena en la cuál se realizó la observación. Los periodos se han codificado de 1 a 5 del siguiente modo:

- Periodo 1: del 16 al 31 de julio
- Periodo 2: del 1 al 15 de agosto

- Periodo 3: del 16 al 31 de agosto
- Periodo 4: del 1 al 15 de septiembre
- Periodo 5: del 16 al 30 de septiembre

LCC: (*Longitud Curva del Caparazón*) Longitud del caparazón desde la cabeza a la cola, medida con una cinta flexible que se adapta a la curvatura del caparazón. Se mide en centímetros.

ACC: (*Anchura Curva del Caparazón*) Anchura en centímetros del caparazón, medida de lado a lado también con una cinta flexible.

Peso: Peso del animal en kilogramos.

Sexo: Las tortugas se muestrearon en playas durante el anidamiento, por lo que la gran mayoría son hembras. No obstante en algunos casos se detectaron machos que también fueron medidos y pesados.

Huevos: Cada hembra se observó en el momento de construir un nido en la arena, en el que realizó una puesta de huevos. En esta variable se registra el tamaño de la puesta (número total de huevos en el nido tras la puesta).

Playa: Nombre de la playa donde se realizó la observación. Se han codificado como A (Ervatao), B (Ponta Cosme), C (Calheta) y D (Porto Ferreira).

Distancia: Distancia (en metros) medida perpendicularmente desde la linea de marea alta hasta la posición del nido.

profNido: Profundidad del nido en centímetros.

Las siguientes tres variables no están disponibles para todos los nidos, sino sólo para aquellos a los que se pudo hacer un seguimiento antes del final de la campaña:

crias_Emerg: número de crías emergidas a la superficie.

crias_Muertas: tortugas encontradas muertas dentro del nido. Son tortugas que, tras haber salido del huevo, no pudieron alcanzar la superficie.

hrotos: diferencia entre el total de la puesta y las crías emergidas más las crías muertas. Corresponde al total de huevos que no produjeron ninguna cría viable: huevos no fecundados, huevos depredados, afectados por microorganismos patógenos, descompuestos por inundaciones causadas por fuertes lluvias, rotos por otras tortugas que anidan en el mismo lugar en que ya había huevos enterrados, ...

cangrejos: variable indicatriz de la posible presencia de cangrejos en el nido. Los cangrejos son los principales depredadores de huevos de tortuga. Esta variable vale 1 si en el nido se han encontrado cangrejos, túneles realizados por cangrejos o cualquier otra señal que delate la presencia de cangrejos. Vale 0 cuando no hay ninguna señal de cangrejos.

Descarga de los datos

Para la realización de esta práctica se utilizarán datos simulados, diferentes para cada alumno. Estos datos pueden obtenerse arrancando R e introduciendo la siguiente instrucción:

```
> source("http://dl.dropbox.com/u/7610774/CursoR/datosTortugas.R")
```

El sistema presentará una pequeña ventana en la que el alumno se identificará mediante su DNI. Tras pulsar en Generar Datos, el archivo con los datos se guardará en la carpeta de usuario (usualmente "*Mis Documentos*") con el nombre tortugas_NNNNNNNN.csv donde NNNNNNNN es el número del DNI proporcionado. Además estos datos quedan ya listos para su uso en el `data.frame` tortugas del entorno R .

Antes de comenzar el tratamiento de los datos, asigna etiquetas a los periodos de observación, a los nombres de las playas y a la presencia/ausencia de cangrejos en los nidos, de forma que los niveles de estos factores queden bien identificados. Para ello utiliza la siguiente sintaxis (córtala y pégala en la ventana de edición de código de R , y ejecútala a continuación):

```
> tortugas$periodo = factor(tortugas$periodo, levels = 1:5,
    labels = c("del 16 al 31 de julio", "del 1 al 15 de agosto",
        "del 16 al 31 de agosto", "del 1 al 15 de septiembre",
        "del 16 al 30 de septiembre"))
> tortugas$playa = factor(tortugas$playa, levels = c("A",
    "B", "C", "D"), labels = c("Ervatao", "Ponta Cosme",
    "Calheta", "Porto Ferreira"))
> tortugas$cangrejos = factor(tortugas$cangrejos, levels = 0:1,
    labels = c("NO", "SÍ"))
> attach(tortugas)
```

Utilizando estos datos, lleva a cabo las tareas y responde a las cuestiones que se enuncian a continuación.

1. **Descripción Morfológica de las tortugas.** Completa la tabla siguiente:

Variable	Media	Mediana	Desv. típica	Mínimo	Máximo	Asimetría
LCC	81.93267	81.7	4.514921	67.4	104.9	0.9196001
ACC	76.52193	76.5	4.533297	56.3	95.5	-0.006786517
Peso	60.74131	60.5	5.162085	42.3	78.4	0.2608019

Solución:

Dado que hay que realizar los mismos cálculos sobre distintas variables, resulta muy cómodo en este caso implementar una función que lleve a cabo dicha tarea. En dicha función incluímos el acceso a la librería `agricolae` que contiene una función para el cálculo de la asimetría (`skewness()`). Asimismo, en todos los casos (salvo en `skewness()` que lo hace por defecto) especificamos la opción `na.rm=TRUE` para que realice los cálculos ignorando los valores perdidos:

```
> describeMorf=function(x){
        require(agricolae)
        descripcion=c(media=mean(x, na.rm=TRUE),
                  mediana=median(x,na.rm=TRUE),
                  desvTip=sd(x,na.rm=TRUE),
                  minimo=min(x,na.rm=TRUE),
                  maximo=max(x,na.rm=TRUE),
                  asim=skewness(x))
        return(descripcion)
 }
```

Aplicamos la función `attach()` para acceder a los datos sin tener que incluir el nombre del data.frame:

```
> attach(tortugas)
```

Ahora aplicamos la función `describeMorf()` a cada una de las variables especificadas:

```
> describeMorf(LCC)
```

```
   media  mediana  desvTip   minimo   maximo     asim
 81.9327  81.7000   4.5149  67.4000 104.9000   0.9196
```

```
> describeMorf(ACC)
```

```
    media   mediana   desvTip    minimo    maximo       asim
76.521929 76.500000  4.533297 56.300000 95.500000 -0.006787
```

```
> describeMorf(peso)
```

```
  media mediana desvTip  minimo  maximo    asim
60.7413 60.5000  5.1621 42.3000 78.4000  0.2608
```

Respuesta correcta:

Variable	Media	Mediana	Desv. típica	Mínimo	Máximo	Asimetría
LCC	81.93	81.7	4.515	67.4	104.9	0.9196
ACC	76.52	76.5	4.533	56.3	95.5	-0.006787
Peso	60.74	60.5	5.162	42.3	78.4	0.2608

Calificación: 1.25 (18/18)

2. **Descripción de los nidos.** Completa la tabla siguiente:

Variable	Media	Mediana	Desv. típica	Mínimo	Máximo	Asimetría
Núm. Huevos	83.79684	83	12.83236	53	141	0.4663074
Distancia	19.94181	19.2	6.571626	3.7	48.5	0.624802
Prof. Nido	54.67268	54.55	5.979197	35.8	78	0.1935636
Crías Emerg.	27.18182	19	26.95177	0	117	0.7020097
Crías Muertas	1.397059	0	2.518758	0	13	2.354128

Solución:

De modo análogo a como hemos hecho en el apartado anterior aplicamos la función `describeMorf()` a las nuevas variables:

```
> describeMorf(Huevos)

   media  mediana  desvTip   minimo   maximo     asim
 83.7968  83.0000  12.8324  53.0000 141.0000   0.4663

> describeMorf(distancia)

  media mediana desvTip  minimo  maximo    asim
19.9418 19.2000  6.5716  3.7000 48.5000  0.6248

> describeMorf(profNido)

  media mediana desvTip  minimo  maximo    asim
54.6727 54.5500  5.9792 35.8000 78.0000  0.1936

> describeMorf(crias_Emerg)
```

```
  media mediana desvTip  minimo  maximo    asim
 27.182  19.000  26.952   0.000 117.000   0.702

> describeMorf(crias_Muertas)

  media mediana desvTip  minimo  maximo    asim
  1.397   0.000   2.519   0.000  13.000   2.354
```

Respuesta correcta:

Variable	Media	Mediana	Desv. típica	Mínimo	Máximo	Asimetría
Núm. Huevos	83.8	83	12.83	53	141	0.4663
Distancia	19.94	19.2	6.572	3.7	48.5	0.6248
Prof. Nido	54.67	54.55	5.979	35.8	78	0.1936
Crías Emerg.	27.18	19	26.95	0	117	0.702
Crías Muertas	1.397	0	2.519	0	13	2.354

Calificación: 1.25 (30/30)

3. **Anidamiento por playas.** Anota a continuación el número de nidos observado en cada playa:

Playa	Ervatao	Ponta Cosme	Calheta	Porto Ferreira
Núm. Nidos	151	166	496	286

Representa estos valores mediante un diagrama de barras y mediante un diagrama de sectores. Copia a continuación la sintaxis que has empleado para ello:

Diagrama de Barras	barplot(table(playa),horiz=FALSE)
Diagrama de sectores	pie(table(playa))

Solución:

Dado que cada nido es construido por una hembra, para calcular el número de nidos por playa bastará con contar el número de hembras registradas en cada playa. Para ello utilizamos la función `table()` para que nos haga un recuento de ejemplares de cada sexo en cada playa:

```
> table(sexo, playa)
```

```
        playa
sexo     Ervatao Ponta Cosme Calheta Porto Ferreira
  Hembra     144         154     453            263
  Macho        7          12      43             23
```

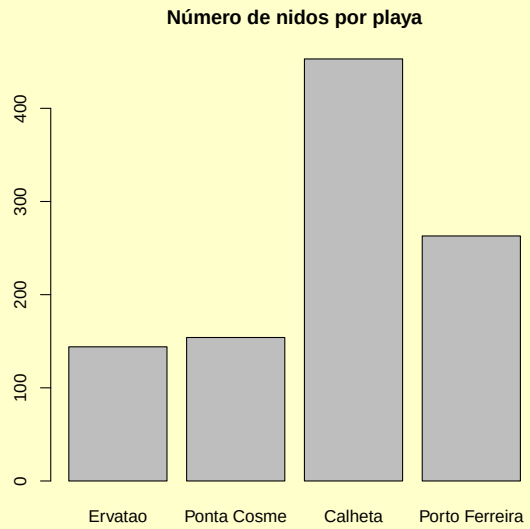
o bien para que nos cuente sólo el número de hembras por playa:

```
> table(playa[sexo == "Hembra"])
```

```
      Ervatao    Ponta Cosme        Calheta Porto Ferreira
          144            154            453            263
```

Para construir un diagrama de barras empleamos la sintaxis siguiente:

```
> barplot(table(playa[sexo == "Hembra"]), main = "Número de nidos por playa")
```



y para el diagrama de sectores:

```
> pie(table(playa[sexo == "Hembra"]), main = "Número de nidos por playa")
```

Respuesta correcta:

Playa	Ervatao	Ponta Cosme	Calheta	Porto Ferreira
Núm. Nidos	144	154	453	263

Diagrama de Barras	`barplot(table(playa[sexo=="Hembra"]))`
Diagrama de sectores	`pie(table(playa[sexo=="Hembra"]))`

Calificación: 0 (0/6)

4. **Diferencias morfológicas entre sexos.** Calcula en primer lugar el número de machos y hembras en tu muestra:

Número de Machos:	85
Número de Hembras:	1014

Calcula ahora el valor medio de cada una de las siguientes variables para cada sexo:

Sexo	LCC	ACC	Peso
Machos	80.81059	69.31412	59.71647
Hembras	82.02673	77.12613	60.82722

Solución:

Podemos contar el número de machos y hembras mediante `table()`:

```
> table(sexo)

sexo
Hembra  Macho
  1014     85
```

Para calcular los valores medios por sexo de las variables solicitadas empleamos `aggregate()`:

```
> aggregate(cbind(LCC, ACC, peso), by = list(Sexo = sexo), mean)

    Sexo   LCC   ACC  peso
1 Hembra 82.03 77.13 60.83
2  Macho 80.81 69.31 59.72
```

Respuesta correcta:

Número de Machos:	85
Número de Hembras:	1014

Medias por sexo:

Sexo	LCC	ACC	Peso
Machos	80.81	69.31	59.72
Hembras	82.03	77.13	60.83

Calificación: 1.25 (8/8)

5. **Frecuencias por rango de talla.** En este apartado nos interesa determinar el número de tortugas dentro de determinados rangos de Longitud Curva del Caparazón. Utilizando el comando `table.freq()` del paquete `agricolae` (ver Tema 0, sección 5.2, página 21 y siguientes), completa la siguiente tabla:

Rango de LCC (cm.)	Frecuencia absoluta	Frecuencia relativa
70-75	43	0.039126479
75-80	299	0.272065514
80-85	571	0.519563239

Solución:

La siguiente sintaxis nos permite obtener las frecuencias pedidas:

```
> require(agricolae)
> table.freq(hist(LCC,br=seq(65,105,by=5),plot=FALSE))
```

```
  Inf Sup    MC  fi     fri   Fi     Fri
   65  70  67.5   5 0.00455    5 0.00455
   70  75  72.5  43 0.03913   48 0.04368
   75  80  77.5 299 0.27207  347 0.31574
   80  85  82.5 571 0.51956  918 0.83530
   85  90  87.5 135 0.12284 1053 0.95814
   90  95  92.5  25 0.02275 1078 0.98089
   95 100  97.5  15 0.01365 1093 0.99454
  100 105 102.5   6 0.00546 1099 1.00000
```

Respuesta correcta:

Rango de LCC (cm.)	Frecuencia absoluta	Frecuencia relativa
70-75	43	0.03913
75-80	299	0.2721
80-85	571	0.5196

Calificación: 1.25 (6/6)

6. **Preferencias por playa según sexo.** Interesa saber si las preferencias por las distintas playas difieren sustancialmente entre sexos. Construye una tabla cruzada playa$\times$sexo que muestre qué proporción de machos se ha encontrado en cada playa. Idem de hembras.

Playa	Proporciones Hembras	Proporciones Machos
Playa de Ervatao:	0.142	0.082
Playa de Ponta Cosme:	0.152	0.141
Playa de Calheta:	0.447	0.506
Playa de Porto Ferreira:	0.259	0.271

Solución:

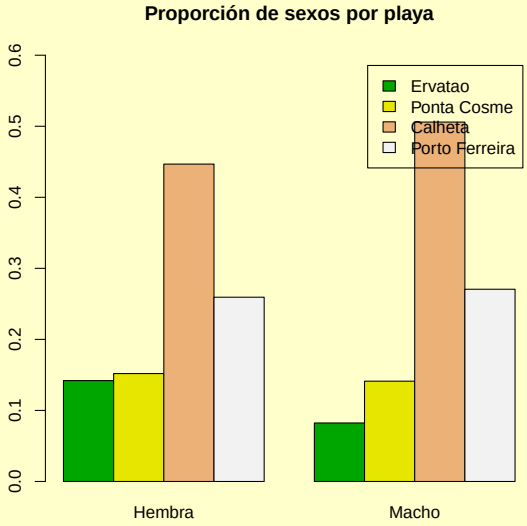
Bastará utilizar la combinación de `table()` y `prop.table()`, indicando que se desea calcular la proporción por columnas:

```
> prop.table(table(playa, sexo), 2)

                sexo
playa             Hembra   Macho
  Ervatao        0.14201 0.08235
  Ponta Cosme    0.15187 0.14118
  Calheta        0.44675 0.50588
  Porto Ferreira 0.25937 0.27059
```

Esta información puede visualizarse en un gráfico como el siguiente:

```
> barplot(prop.table(table(playa, sexo), 2), beside = T,
    legend.text = T, main = "Proporción de sexos por playa",
    ylim = c(0, 0.6), col = terrain.colors(4))
```



Se aprecia que las proporciones de machos y hembras son muy parecidas en todas las playas, por lo que a priori no parece haber indicios fuertes de distintas preferencias por playa según sexo.

Respuesta correcta:

Playa	Proporciones Hembras	Proporciones Machos
Playa de Ervatao:	0.142	0.08235
Playa de Ponta Cosme:	0.1519	0.1412
Playa de Calheta:	0.4467	0.5059
Playa de Porto Ferreira:	0.2594	0.2706

Calificación: 1.25 (8/8)

7. **Tamaño de la puesta según periodo.** Otra cuestión de interés consiste en saber si los tamaños de las puestas (nº de huevos en el nido) difieren de manera importante según los periodos de observación. Anota a continuación el tamaño medio de la puesta, así como el primer y tercer cuartil de dicho tamaño, en cada periodo:

Periodo	Número medio de Huevos por nido	Primer Cuartil	Tercer Cuartil
Del 16 al 31 de julio	88.00552	79	96
Del 1 al 15 de agosto	89.04977	81	96
Del 16 al 31 de agosto	83.97596	76.75	91.00
Del 1 al 15 de septiembre	80.125	72	86
Del 16 al 30 de septiembre	77.69388	70	85

Representa gráficamente estos datos mediante un diagrama de caja y bigotes (boxplot). Introduce a continuación la sintaxis que has empleado para ello:

```
boxplot(by(Huevos,periodo,quantile,na,rm=TRUE,probs=c(0,25,0,75)))
```

Otra posible cuestión a tener en cuenta en la exploración de estos datos es si la variación en el tamaño de las puestas en los distintos periodos ha sido la misma durante todos los años de observación (1999 a 2004). La siguiente sintaxis (no descrita en el tema 0) permite construir un panel de boxplots que permite visualizar esta información:

```
library(lattice) # Se requiere cargar esta librería

bwplot(Huevos~periodo|Anno,data=tortugas)
```

Solución:

Construimos en primer lugar una función `resumen()` que calcule los parámetros de interés (media y primer y tercer cuartil):

```
> resumen = function(x) {
     params = c(media = mean(x, na.rm = TRUE), quantile(x,
         probs = c(0.25, 0.75), na.rm = TRUE))
     return(params)
 }
```

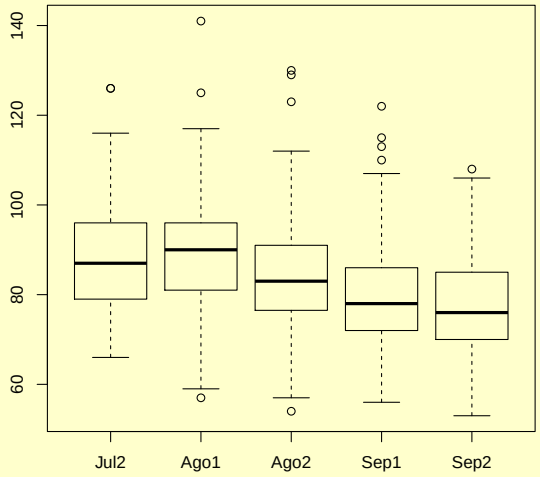
Ahora aplicamos `aggregate()` para calcular estos parámetros según periodo:

```
> aggregate(Huevos, by = list(Periodo = periodo), resumen)
```

```
                    Periodo x.media x.25% x.75%
1      del 16 al 31 de julio   88.01 79.00 96.00
2      del 1 al 15 de agosto   89.05 81.00 96.00
3     del 16 al 31 de agosto   83.98 76.75 91.00
4  del 1 al 15 de septiembre   80.12 72.00 86.00
5 del 16 al 30 de septiembre   77.69 70.00 85.00
```

Podemos construir un boxplot de manera sencilla mediante:

```
> boxplot(Huevos ~ periodo, names = c("Jul2", "Ago1", "Ago2",
     "Sep1", "Sep2"), cex.lab = 0.7)
```

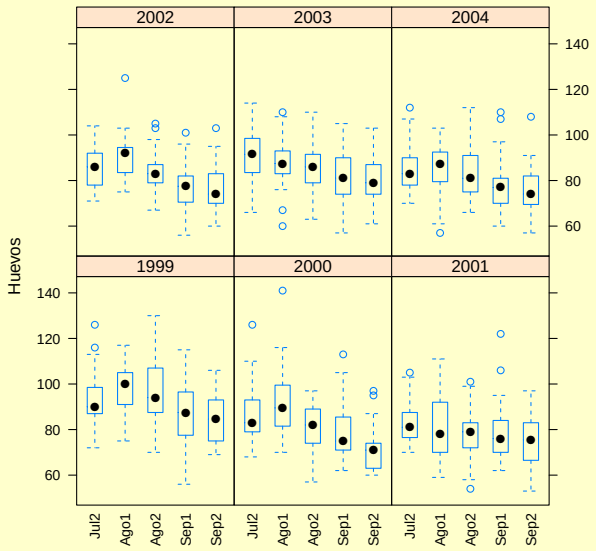


Hemos recodificado los nombres de los periodos de observación en función del mes y la quincena, para que se puedan visualizar en la base de cada boxplot.

Como se señala en el texto, si se desea visualizar si el tamaño de las puestas por periodo presenta variación interanual podemos construir el gráfico siguiente (se ha hecho alguna modificación menor para las etiquetas de los ejes):

```
> require(lattice)
> bwplot(Huevos~periodo|factor(Anno),
```

```
scales = list(x = list(labels=c("Jul2","Ago1","Ago2","Sep1","Sep2"),
                       rot=90)))
```



Respuesta correcta:

Periodo	Número medio de Huevos por nido	Primer Cuartil	Tercer Cuartil
Del 16 al 31 de julio	88.01	79	96
Del 1 al 15 de agosto	89.05	81	96
Del 16 al 31 de agosto	83.98	76.75	91
Del 1 al 15 de septiembre	80.12	72	86
Del 16 al 30 de septiembre	77.69	70	85

Sintaxis para el diagrama de caja y bigote

```
boxplot(Huevos~periodo)
```

Calificación: 1.1719 (15/16)

8. **Asociación entre las medidas longitudinal y transversal del caparazón.** Construye un gráfico de nube de puntos representando ACC en el eje Y frente a LCC en el eje X. Dibuja de color distinto los puntos correspondientes a machos y hembras y representa la linea de regresión para cada sexo.

   a) ¿Sugiere el gráfico que la relación LCC-ACC en machos pueda ser diferente que en hembras?

Sí

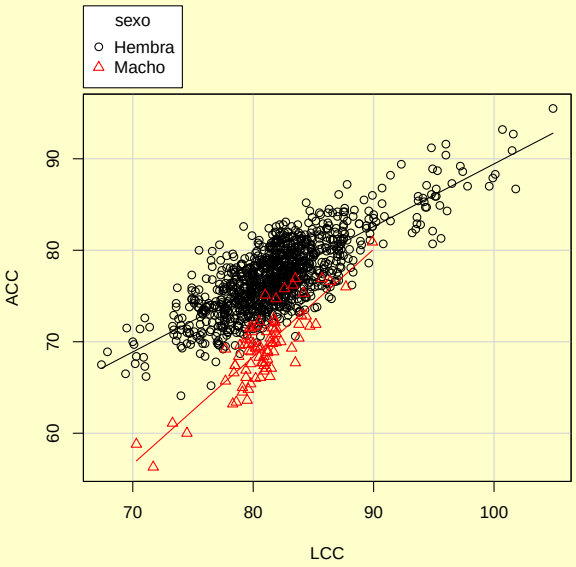
b) Completa la siguiente tabla con los coeficientes de la regresión de ACC frente a LCC y con los coeficientes de correlación entre ambas variables para cada sexo:

Sexo	Ordenada (Intercept)	Pendiente	Correlación
Machos	-25.541	1.174	0.8284804
Hembras	20.9249	0.6852	0.7860829

Solución:

Para hacer la gráfica de la recta de regresión LCC-ACC para cada sexo, ejecutamos:

```
> require(car)
> scatterplot(ACC ~ LCC | sexo, smooth = F)
```



El gráfico sugiere que la recta para los machos tiene una pendiente más acusada que para las hembras, lo que revelaría diferencias morfológicas entre ambos sexos. En la gráfica se observa que cuando LCC está entre 75 y 85 cm los machos tienen una ACC más corta que las hembras (por tanto presentarían una forma más alargada). Para tallas mayores (entre 95 y 100 cm) la morfología de ambos sexos sería similar.

Para calcular los coeficientes de regresión, así como las correlaciones entre ambas variables, utilizamos `lm()` y `cor` respectivamente:

- Para los machos:

```
> lm(ACC ~ LCC, data = subset(tortugas, sexo == "Macho"))

Call:
lm(formula = ACC ~ LCC, data = subset(tortugas, sexo == "Macho"))
```

```
Coefficients:
(Intercept)          LCC
     -25.54         1.17

> cor(ACC[sexo == "Macho"], LCC[sexo == "Macho"])

[1] 0.8285
```

-

```
> lm(ACC ~ LCC, data = subset(tortugas, sexo == "Hembra"))

Call:
lm(formula = ACC ~ LCC, data = subset(tortugas, sexo == "Hembra"))

Coefficients:
(Intercept)          LCC
     20.925        0.685

> cor(ACC[sexo == "Hembra"], LCC[sexo == "Hembra"])

[1] 0.7861
```

Respuesta correcta:

a) ¿Sugiere el gráfico que la relación LCC-ACC en machos pueda ser diferente que en hembras?

Sí

b) Coeficientes de la regresión de ACC frente a LCC y coeficientes de correlación entre ambas variables para cada sexo:

Sexo	Ordenada (Intercept)	Pendiente	Correlación
Machos	-25.54	1.174	0.8285
Hembras	20.92	0.6852	0.7861

Calificación: 1.25 (7/7)