

Facultad de Ciencias del Mar. Curso 2010-11

Estadística – Práctica nº 5.

Descarga el archivo [sargos.csv](http://dl.dropbox.com/u/7610774/sargos.csv) de la dirección <http://dl.dropbox.com/u/7610774/sargos.csv> y léelo con R. Puedes leerlo directamente desde la red mediante el comando:

```
sargos = read.table(file = "http://dl.dropbox.com/u/7610774/sargos.csv",
  sep = ";", dec = ",", header = TRUE)
attach(sargos)
```

1. Utilizando el comando `t.test()` :

a) Calcula intervalos de confianza al 95 % para la media de las variables `ldors`, `lpect`, `loper`, `altop`.

Solución:

El comando `t.test()` por defecto no sólo calcula el intervalo de confianza, sino que además contrasta la hipótesis de si la media de la variable que se indica es 0 (hipótesis nula) o no (hipótesis alternativa):

```
t.test(ldors)

##
## One Sample t-test
##
## data: ldors
## t = 74.24, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.938 5.207
## sample estimates:
## mean of x
##      5.073
##
```

La función `str()` nos muestra la estructura del objeto creado con `t.test()`. Observemos las variables contenidas en dicho objeto:

```
str(t.test(ldors))

## List of 9
## $ statistic : Named num 74.2
## .. attr(*, "names")= chr "t"
## $ parameter : Named num 199
## .. attr(*, "names")= chr "df"
## $ p.value : num 5.09e-147
## $ conf.int : atomic [1:2] 4.94 5.21
## .. attr(*, "conf.level")= num 0.95
## $ estimate : Named num 5.07
## .. attr(*, "names")= chr "mean of x"
## $ null.value : Named num 0
## .. attr(*, "names")= chr "mean"
## $ alternative: chr "two.sided"
## $ method : chr "One Sample t-test"
## $ data.name : chr "ldors"
## - attr(*, "class")= chr "htest"
```

Tal como vemos, el intervalo de confianza para la media de *ldors* se devuelve en la variable "conf.int". De esta forma, si sólo estamos interesados en mostrar dicho intervalo, nos basta con ejecutar:

```
c(t.test(ldors)$conf.int)
## [1] 4.938 5.207

c(t.test(lpect)$conf.int)
## [1] 4.739 5.014

c(t.test(loper)$conf.int)
## [1] 4.254 4.510

c(t.test(altop)$conf.int)
## [1] 6.595 6.937
```

b) Contrastar, con un nivel de significación del 5% si los valores medios de las siguientes variables difieren entre machos y hembras: *ldors*, *lpect*, *loper*, *altop* y *phig*.

Solución:

En todos los casos se trata de realizar contrastes de la forma:

$$\begin{cases} H_0 : \mu_M = \mu_H \\ H_0 : \mu_M \neq \mu_H \end{cases}$$

Si suponemos que las variables son normales, tales contrastes se pueden llevar a cabo mediante `t.test()` :

```
t.test(ldors ~ sexo)

##
## Welch Two Sample t-test
##
## data: ldors by sexo
## t = 5.647, df = 174, p-value = 6.537e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.4769 0.9894
## sample estimates:
## mean in group Hembra mean in group Macho
## 5.392 4.659
##
```

```
t.test(lpect ~ sexo)
```

```
##  
## Welch Two Sample t-test  
##  
## data: lpect by sexo  
## t = -1.711, df = 184.6, p-value = 0.08882  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.51478 0.03665  
## sample estimates:  
## mean in group Hembra mean in group Macho  
## 4.773 5.012  
##
```

```
t.test(loper ~ sexo)
```

```
##  
## Welch Two Sample t-test  
##  
## data: loper by sexo  
## t = -2.145, df = 174, p-value = 0.03333  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.54152 -0.02253  
## sample estimates:  
## mean in group Hembra mean in group Macho  
## 4.259 4.541  
##
```

```
t.test(altop ~ sexo)
```

```
##  
## Welch Two Sample t-test  
##  
## data: altop by sexo  
## t = -1.771, df = 174.8, p-value = 0.07824  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.66110 0.03571  
## sample estimates:  
## mean in group Hembra mean in group Macho  
## 6.630 6.943  
##
```

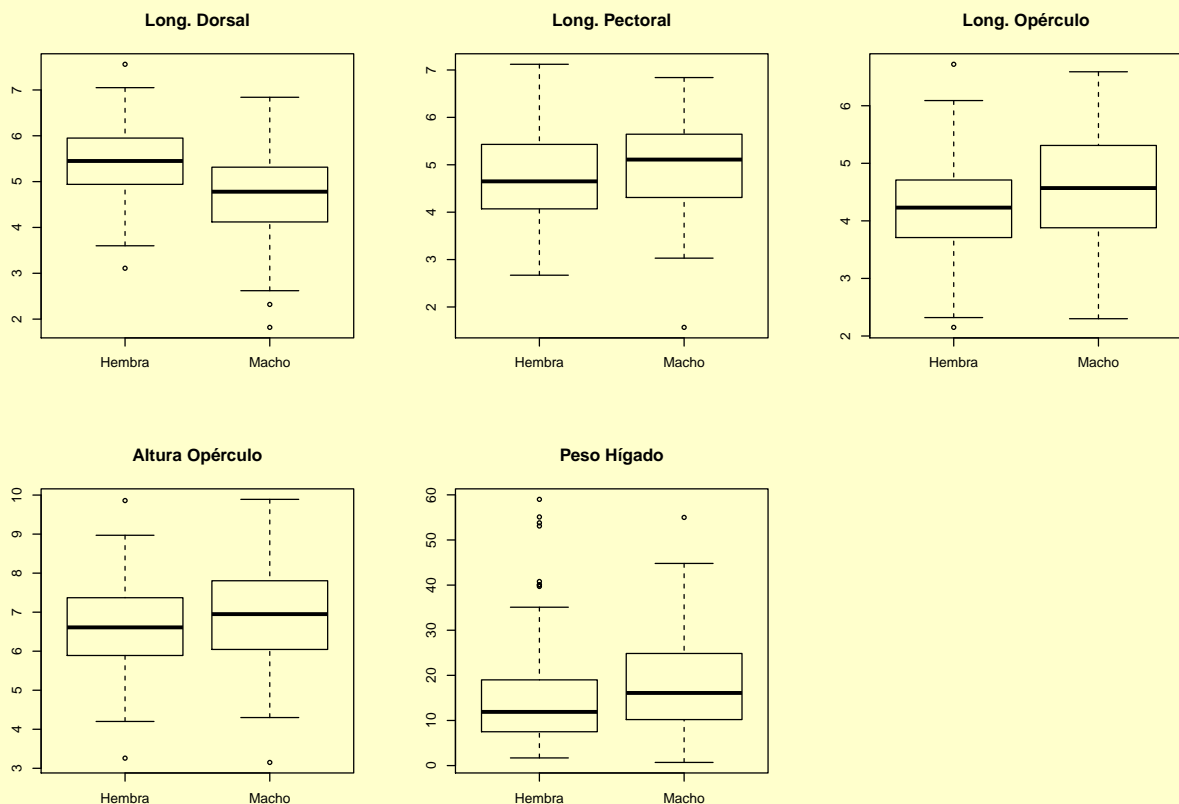
```
t.test(phig ~ sexo)
```

```
##  
## Welch Two Sample t-test  
##  
## data: phig by sexo  
## t = -1.723, df = 193.5, p-value = 0.08651  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -5.7883 0.3907  
## sample estimates:  
## mean in group Hembra mean in group Macho  
## 15.36 18.06  
##
```

Atendiendo a los p-valores, observamos que sólo en los casos de *ldors* ($p=6.537e-08$) y *loper* ($p=0.03333$) el p-valor es inferior a $\alpha = 0,05$, por lo que podemos concluir que ambas variables difieren en promedio de manera significativa entre machos y hembras. En el resto de variables no se detectan diferencias entre sexos.

No obstante, este procedimiento no es del todo correcto toda vez que no hemos explorado los datos para comprobar si se verifica la hipótesis de normalidad. En principio la siguiente figura nos muestra gráficamente la distribución de estas variables en ambos sexos:

```
par(mfrow = c(2, 3))
boxplot(ldors ~ sexo, main = "Long. Dorsal")
boxplot(lppect ~ sexo, main = "Long. Pectoral")
boxplot(loper ~ sexo, main = "Long. Opérculo")
boxplot(altop ~ sexo, main = "Altura Opérculo")
boxplot(phig ~ sexo, main = "Peso Hígado")
```



En el caso del peso del hígado se aprecia que la distribución de esta variable en ambos sexos es asimétrica, lo que indica una previsible falta de normalidad. Contrastamos la hipótesis de normalidad:

$$\begin{cases} H_0 : X \approx N(\mu, \sigma) \\ H_1 : X \not\approx N(\mu, \sigma) \end{cases}$$

mediante el test de Shapiro-Wilk:

```
shapiro.test(ldors[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: ldors[sexo == "Macho"]  
## W = 0.9883, p-value = 0.6324  
##
```

```
shapiro.test(ldors[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: ldors[sexo == "Hembra"]  
## W = 0.9898, p-value = 0.5625  
##
```

```
shapiro.test(lpect[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: lpect[sexo == "Macho"]  
## W = 0.9791, p-value = 0.1708  
##
```

```
shapiro.test(lpect[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: lpect[sexo == "Hembra"]  
## W = 0.9862, p-value = 0.2998  
##
```

```
shapiro.test(loper[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: loper[sexo == "Macho"]  
## W = 0.9787, p-value = 0.1624  
##
```

```
shapiro.test(loper[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: loper[sexo == "Hembra"]  
## W = 0.9887, p-value = 0.4696  
##
```

```
shapiro.test(altop[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  altop[sexo == "Macho"]  
## W = 0.9948, p-value = 0.9822  
##
```

```
shapiro.test(altop[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  altop[sexo == "Hembra"]  
## W = 0.9959, p-value = 0.9848  
##
```

```
shapiro.test(phig[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  phig[sexo == "Macho"]  
## W = 0.9488, p-value = 0.001751  
##
```

```
shapiro.test(phig[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  phig[sexo == "Hembra"]  
## W = 0.8141, p-value = 1.246e-10  
##
```

Tal como preveíamos, puede aceptarse la normalidad (p-valores mayores que 0.05) para todas las variables en ambos sexos salvo para el peso del hígado (en ambos sexos el p-valor del test de Shapiro es menor que 0.05). Por tanto la comparación de medias entre sexos para esta última variable utilizando el **t.test()** es incorrecta, y en su lugar hemos de utilizar el test de Wilcoxon:

```
wilcox.test(phig ~ sexo)
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data:  phig by sexo  
## W = 3893, p-value = 0.01178  
## alternative hypothesis: true location shift is not equal to 0  
##
```

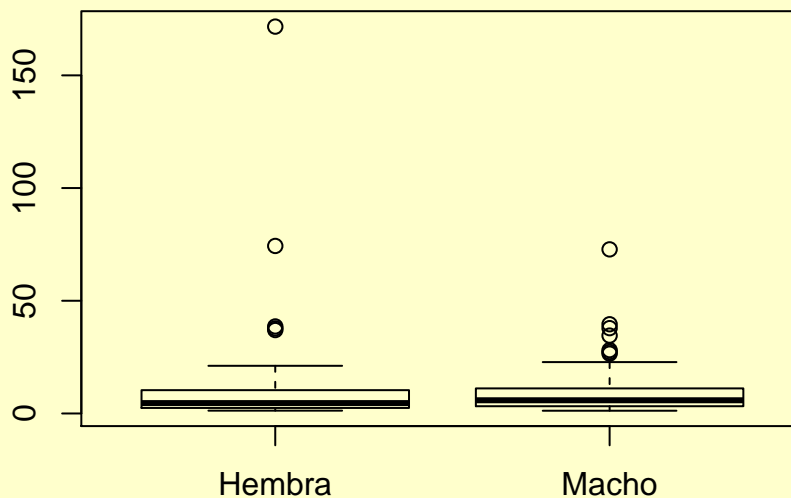
El p-valor de este contraste (0.01178) indica que existen diferencias significativas en el peso del hígado entre machos y hembras: la gráfica anterior nos permite observar que dichas diferencias obedecen a que las hembras tienden, en general, a tener un hígado con menos peso que los machos.

- c) ¿Puede asegurarse, con un 5% de significación, que el peso medio de las gónadas de las hembras es superior al peso medio de las gónadas de los machos?

Solución:

Comprobamos en primer lugar la normalidad de la variable:

```
boxplot (pgon ~ sexo)
```



El gráfico nos indica la presencia de una fuerte asimetría en los datos para cada sexo, por lo que es previsible que no haya normalidad. Llevamos a cabo el test de Shapiro-Wilk:

```
shapiro.test (pgon[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  pgon[sexo == "Macho"]  
## W = 0.6734, p-value = 6.918e-10  
##
```

```
shapiro.test (pgon[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  pgon[sexo == "Hembra"]  
## W = 0.4093, p-value = 2.323e-12  
##
```

Los p-valores pequeños muestran claramente que el peso de las gónadas no es normal, por lo que realizamos la comparación de las tendencias centrales a través del test de Wilcoxon; la hipótesis alternativa a considerar es $H_1 : \mu_H > \mu_M$:

```
wilcox.test(pgon ~ sexo, alternative = "greater")

##
## Wilcoxon rank sum test with continuity correction
##
## data:  pgon by sexo
## W = 1144, p-value = 0.8336
## alternative hypothesis: true location shift is greater than 0
##
```

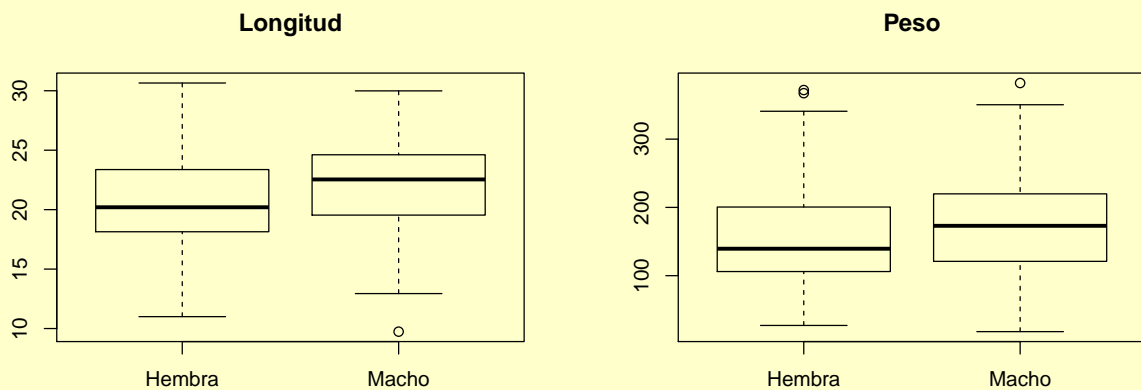
El p-valor obtenido (0.8336) indica que no hay evidencia suficiente de que el peso de las gónadas en hembras sea superior al peso de las gónadas en machos. Observando la figura podemos comprobar que, en general, salvo para una hembra cuyas gónadas pesan más de 150 mg, en el resto de sujetos, sin distinguir sexos, la distribución del peso de las gónadas es muy similar. Ello explica que el test de Wilcoxon no haya detectado diferencias, ya que este test es "insensible" a la presencia de valores extremos como el señalado en el caso de las hembras.

d) ¿Puede asegurarse con un 5% de significación que el peso y la longitud medios son mayores en los machos que en las hembras?

Solución:

Verificamos primero la normalidad de ambas variables en cada sexo:

```
par(mfrow = c(1, 2))
boxplot(long ~ sexo, main = "Longitud")
boxplot(peso ~ sexo, main = "Peso")
```




```
shapiro.test(long[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: long[sexo == "Macho"]  
## W = 0.984, p-value = 0.36  
##
```

```
shapiro.test(long[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: long[sexo == "Hembra"]  
## W = 0.9866, p-value = 0.3218  
##
```

```
shapiro.test(peso[sexo == "Macho"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: peso[sexo == "Macho"]  
## W = 0.9817, p-value = 0.2577  
##
```

```
shapiro.test(peso[sexo == "Hembra"])
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: peso[sexo == "Hembra"]  
## W = 0.9406, p-value = 7.9e-05  
##
```

Como vemos, puede aceptarse la normalidad de la longitud (p-valor del test de Shapiro >0.05) en ambos sexos, pero no así del peso: es normal para los machos (p-valor=0.257), pero no para las hembras (p-valor<0.001).

Contrastamos entonces si la longitud es en promedio mayor en machos que en hembras mediante el **t.test()** ; la hipótesis alternativa es, pues, $H_1 : \mu_H < \mu_M$ (téngase en cuenta que en R la ordenación de valores para variables alfanuméricas sigue el orden alfabético; por tanto a la hora de definir el contraste las hembras van en primer lugar y los machos en segundo):

```
t.test(long ~ sexo, alternative = "less")
```

```
##  
## Welch Two Sample t-test  
##  
## data: long by sexo  
## t = -2.101, df = 181.5, p-value = 0.01853  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
##      -Inf -0.2472  
## sample estimates:  
## mean in group Hembra mean in group Macho  
##           20.84           22.00  
##
```

Podemos concluir que existe evidencia suficiente (p-valor 0.018) de que los machos miden más que las hembras; la diferencia estimada es de $22 - 20,84 = 1,16$ centímetros, pudiendo asegurar con un 95 % de confianza que la diferencia es de al menos 0,247 cm.

En cuanto al peso, ante la ausencia de normalidad podemos aplicar el test de Wilcoxon:

```
wilcox.test(peso ~ sexo, alternative = "less")  
  
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data: peso by sexo  
## W = 4012, p-value = 0.01299  
## alternative hypothesis: true location shift is less than 0  
##
```

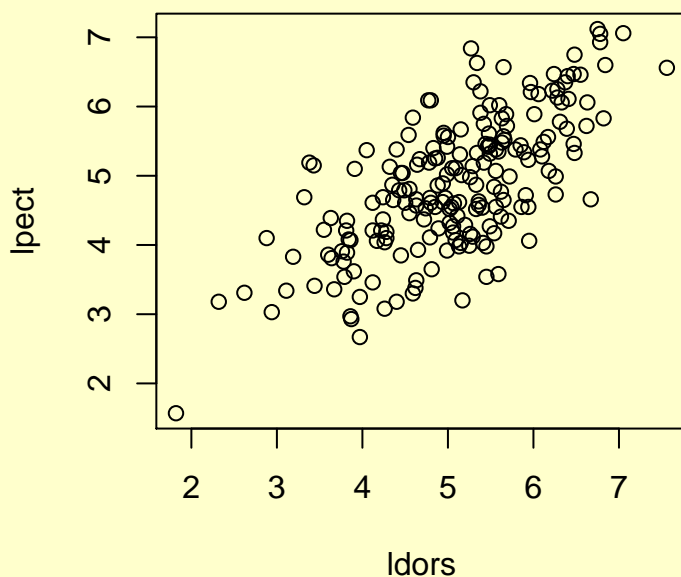
que nos indica que también existe evidencia suficiente de que el peso medio de los machos tiende a ser superior al de las hembras (p-valor 0.013).

e) ¿Es significativa al 5% la diferencia entre los valores medios de `ldors` y `lpect` para cada pez?

Solución:

Para cada pez cabe suponer que ambas variables estén fuertemente relacionadas. Podemos comprobarlo visualmente en el siguiente gráfico:

```
par(mfrow = c(1, 1))  
plot(ldors, lpect)
```



La media de cada una de estas variables es:

```

mean(ldors)

## [1] 5.073

mean(lpect)

## [1] 4.877

```

Como vemos, son valores parecidos; para determinar si la diferencia es significativa debemos llevar a cabo un contraste de medias emparejadas. Como ya hemos comprobado que estas variables son normales podemos proceder directamente con un `t.test()`. Además hacemos el contraste por separado para hembras y machos pues ya hemos detectado diferencias entre sexos en la variable `ldors`:

```

t.test(ldors[sexo == "Macho"], lpect[sexo == "Macho"], paired = T)

##
## Paired t-test
##
## data: ldors[sexo == "Macho"] and lpect[sexo == "Macho"]
## t = -5.813, df = 86, p-value = 1.016e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.4740 -0.2324
## sample estimates:
## mean of the differences
## -0.3532
##

t.test(ldors[sexo == "Hembra"], lpect[sexo == "Hembra"], paired = T)

##
## Paired t-test
##
## data: ldors[sexo == "Hembra"] and lpect[sexo == "Hembra"]
## t = 10.89, df = 112, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.5064 0.7316
## sample estimates:
## mean of the differences
## 0.619
##

```

En ambos casos existen diferencias significativas en los valores promedio de ambas variables (p-valores respectivos menores que 0.001). La magnitud de la diferencia es negativa en machos (por término medio `ldors` es 0.352 cm menor que `lpect`), mientras que es positiva en hembras (`ldors` es, en promedio, 0.619 cm. mayor que `lpect`).

2. Utilizando el comando `var.test()` :

- a) Calcula intervalos de confianza al 95 % para las varianzas de las variables `ldors`, `lpect`, `loper`, `altop`.

Solución:

La estimación de las varianzas de las variables indicadas es:

```

var(ldors)
## [1] 0.9338
var(lpect)
## [1] 0.9678
var(loper)
## [1] 0.8411
var(altop)
## [1] 1.508

```

De modo análogo a como ya hicimos para las medias, podemos calcular el intervalo de confianza para las varianzas utilizando `var.test()`, extrayendo sólo el intervalo:

```

library(TeachingDemos)
c(sigma.test(ldors)$conf.int)
## [1] 0.7744 1.1483
c(sigma.test(lpect)$conf.int)
## [1] 0.8026 1.1901
c(sigma.test(loper)$conf.int)
## [1] 0.6976 1.0343
c(sigma.test(altop)$conf.int)
## [1] 1.250 1.854

```

- b) ¿Puede asegurarse, con un 5% de significación, que el peso del hígado presenta más variabilidad en los machos que en las hembras?

Solución:

Podemos estimar la variabilidad en cada sexo mediante la varianza:

```

var(phig[sexo == "Macho"])
## [1] 108.9
var(phig[sexo == "Hembra"])
## [1] 135.9

```

Dado que la varianza en los machos es menor que en las hembras, ni siquiera procede llevar a cabo un contraste de hipótesis, ya que estos datos no muestran ninguna evidencia de que los machos presenten mayor variabilidad que las hembras en esta variable.

Si estuviésemos, no obstante, interesados en la pregunta contraria (¿existe evidencia suficiente de que las hembras presenten más variabilidad que los machos en el peso del hígado?), esto es, en realizar el contraste:

$$\begin{cases} H_0 : \sigma_H^2 \leq \sigma_M^2 \\ H_1 : \sigma_H^2 > \sigma_M^2 \end{cases}$$

puesto que, como ya hemos visto el peso del hígado no sigue una distribución normal para ninguno de los dos sexos, resulta inapropiado el uso del siguiente contraste basado en la chi-cuadrado, ya que éste depende de la hipótesis de normalidad:

```
var.test(phig ~ sexo, alternative = "greater")

##
## F test to compare two variances
##
## data:  phig by sexo
## F = 1.248, num df = 112, denom df = 86, p-value = 0.1415
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.8886      Inf
## sample estimates:
## ratio of variances
##           1.248
##
```

En su lugar, y dado que disponemos de una muestra grande, utilizaremos (véase por ejemplo http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Coeurjolly~et~al.pdf) la función `asympt.test()` incluida en el paquete `asymptTest`, que nos permite realizar un contraste para la diferencia de varianzas en poblaciones no normales con muestras grandes:

```
library(asymptTest)
asympt.test(phig ~ sexo, parameter = "dVar", alternative = "greater")

##
## Two-sample asymptotic difference of variances test
##
## data:  phig by sexo
## statistic = 0.7736, p-value = 0.2196
## alternative hypothesis: true difference of variances is greater than 0
## 95 percent confidence interval:
##  -30.38      Inf
## sample estimates:
## difference of variances
##           26.98
##
```

El p-valor de contraste (0.2196) indica que no existen evidencia suficiente para asegurar que, en general, la variabilidad del peso del hígado sea superior en las hembras que en los machos.

- c) ¿Existe evidencia, al 5% de significación, de que el peso de las gónadas es más variable en las hembras que en los machos?

Solución:

Estimamos en primer lugar las varianzas respectivas (utilizamos la opción `na.rm=TRUE` ya que el peso de las gónadas presenta bastantes valores perdidos):

```
var(pgon[sexo == "Macho"], na.rm = T)
## [1] 163.4
var(pgon[sexo == "Hembra"], na.rm = T)
## [1] 753.6
```

El peso de las gónadas tampoco es normal, por lo que utilizaremos otra vez la función `asympt.test()` :

```
asympt.test(pgon ~ sexo, parameter = "dVar")
##
## Two-sample asymptotic difference of variances test
##
## data:  pgon by sexo
## statistic = 1.062, p-value = 0.2881
## alternative hypothesis: true difference of variances is not equal to 0
## 95 percent confidence interval:
## -498.9 1679.4
## sample estimates:
## difference of variances
## 590.2
##
```

El p-valor de contraste (0.2881) indica que tampoco en este caso se detectan diferencias significativas en variabilidad entre machos y hembras.

d) ¿Existen diferencias significativas al 5% entre machos y hembras en la variabilidad de la posición de las aletas dorsal y pectoral respecto al morro?

Solución:

Estas variables sí siguen distribuciones normales y por tanto podemos calcular el contraste para el cociente de varianzas:

```
var.test(ldors ~ sexo)
##
## F test to compare two variances
##
## data:  ldors by sexo
## F = 0.8012, num df = 112, denom df = 86, p-value = 0.2698
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.5343 1.1886
## sample estimates:
## ratio of variances
## 0.8012
##
```

```

var.test(lpect ~ sexo)

##
## F test to compare two variances
##
## data:  lpect by sexo
## F = 0.9879, num df = 112, denom df = 86, p-value = 0.9451
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6588 1.4656
## sample estimates:
## ratio of variances
##          0.9879
##

```

Ambos p-valores (0.2698 y 0.9451) superan el valor de $\alpha=0.05$, por lo que no hay evidencia suficiente para asegurar que la variabilidad en estas variables sea diferente entre ambos sexos.

3. Utilizando el comando `shapiro.test()` decide si es posible aceptar la normalidad de las variables `ldors`, `lpect`, `loper`, `altop` globalmente y en cada sexo por separado.

Solución:

Ya resuelto en las preguntas anteriores.

4. ¿Puede aceptarse, con una significación del 5 % que el peso medio de las gónadas en los machos en la isla de Gran Canaria es de 14 gramos?. ¿Puede aceptarse la misma hipótesis para las hembras? ¿es significativa la diferencia en el peso medio de las gónadas entre machos y hembras en esta isla?

Solución:

Si la variable `pgon` fuese normal, el contraste podría realizarse directamente mediante:

```

t.test(pgon[sexo == "Macho" & isla == "GC"], mu = 14)

##
## One Sample t-test
##
## data:  pgon[sexo == "Macho" & isla == "GC"]
## t = 0.2487, df = 12, p-value = 0.8078
## alternative hypothesis: true mean is not equal to 14
## 95 percent confidence interval:
##  3.006 27.828
## sample estimates:
## mean of x
##      15.42
##

```

Sin embargo, ya sabemos que el peso de las gónadas no sigue una distribución normal, por lo que este contraste no resulta adecuado. Una alternativa razonable podría ser realizar el test sobre la mediana y no sobre la media, ya que la mediana es también una medida de tendencia central. La mediana, por definición, es aquel valor tal que la mitad de los valores de la variables son mayores o iguales que él, y la otra mitad son menores o iguales. Para contrastar si puede aceptarse que la mediana es 14, podemos llamar $\pi =$ "Proporción de machos cuyas gónadas pesan al menos 14 gramos" y llevar a cabo el contraste:

$$\begin{cases} H_0 : \pi = 0,5 \\ H_1 : \pi \neq 0,5 \end{cases}$$

Para ello definimos primero una variable que especifica si las gónadas pesan más o menos de 14 gramos, tal como se muestra a continuación:

```
pg14 = ifelse(pgon < 14, "Menos de 14 gr.", "14 gr. o más")
table(pg14[sexo == "Macho" & isla == "GC"])

##
##      14 gr. o más Menos de 14 gr.
##              3              10

binom.test(table(pg14[sexo == "Macho" & isla == "GC"]))

##
## Exact binomial test
##
## data:  table(pg14[sexo == "Macho" & isla == "GC"])
## number of successes = 3, number of trials = 13, p-value =
## 0.09229
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.05038 0.53813
## sample estimates:
## probability of success
##              0.2308
##
```

El p-valor (0.09229) indica que no existe evidencia suficiente, al 5% de significación de que el peso mediano de las gónadas de estos machos difiera de 14 gramos, por lo que aceptamos la hipótesis de que $Me = 14$.

Para las hembras el procedimiento sería análogo:

```
table(pg14[sexo == "Hembra" & isla == "GC"])

##
##      14 gr. o más Menos de 14 gr.
##              2              8
```



```
binom.test(table(pgl4[sexo == "Hembra" & isla == "GC"]))

##
## Exact binomial test
##
## data: table(pgl4[sexo == "Hembra" & isla == "GC"])
## number of successes = 2, number of trials = 10, p-value =
## 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
## 0.02521 0.55610
## sample estimates:
## probability of success
## 0.2
##
```

y tampoco hay evidencia suficiente de que el peso mediano de las gónadas en las hembras difiera de 14 gramos (p-valor 0.1094). Señalemos, no obstante, tanto en este caso como en el anterior que los pequeños tamaños muestrales (13 y 12 sujetos respectivamente) implican que los contrastes tienen muy poca potencia, lo que significa una probabilidad importante de aceptar una hipótesis nula falsa.

En cuanto a la comparación del peso medio de las gónadas de machos y hembras en esta isla, utilizamos el test de Wilcoxon:

```
wilcox.test(pgon[isla == "GC"] ~ sexo[isla == "GC"])

##
## Wilcoxon rank sum test
##
## data: pgon[isla == "GC"] by sexo[isla == "GC"]
## W = 53, p-value = 0.4833
## alternative hypothesis: true location shift is not equal to 0
##
```

No detectamos diferencias significativas entre sexos (p-valor 0.4833)

5. ¿Puede aceptarse la hipótesis de que el peso medio de los machos de la isla de Tenerife supera los 180 gramos? ¿Puede aceptarse esta hipótesis para los de Gran Canaria?

Solución:

Ya hemos visto que la variable *peso* es normal para ambos sexos. Por tanto para contrastar estas hipótesis respecto a la media podemos utilizar el `t.test()` :

```
mean(peso[isla == "TF"])

## [1] 175.5

mean(peso[isla == "GC"])

## [1] 166.4
```

```
t.test(peso[isla == "TF"], mu = 180)

##
## One Sample t-test
##
## data: peso[isla == "TF"]
## t = -0.3536, df = 39, p-value = 0.7256
## alternative hypothesis: true mean is not equal to 180
## 95 percent confidence interval:
##  149.7 201.3
## sample estimates:
## mean of x
##    175.5
##
```

```
t.test(peso[isla == "GC"], mu = 180)

##
## One Sample t-test
##
## data: peso[isla == "GC"]
## t = -1.142, df = 47, p-value = 0.2594
## alternative hypothesis: true mean is not equal to 180
## 95 percent confidence interval:
##  142.5 190.4
## sample estimates:
## mean of x
##    166.4
##
```

En ambos casos los p-valores superan el nivel de significación $\alpha = 0,05$ (p-valor=0.7256 para datos de Tenerife, 0.2594 para datos de Gran Canaria), por lo que aunque las medias muestrales son ambas menores que 180 gramos, no constituyen evidencia suficiente para asegurar que el peso medio en ambas poblaciones difiera de dicha cantidad.

6. Utilizando el comando `binom.test()`, y suponiendo que la muestra de sargos es aleatoria, ¿muestran los datos evidencia significativa al 5% de que la proporción de machos es inferior a la proporción de hembras? (**Nota:** téngase en cuenta que la hipótesis de que *la proporción de machos es inferior a la de las hembras* es equivalente a la hipótesis *la proporción de machos es inferior al 50%*). Calcula un intervalo de confianza al 95% para las proporciones respectivas de machos y hembras.

Solución:

Aplicamos `binom.test()` a la variable `sexo`:

```

table(sexo)

## sexo
## Hembra Macho
##      113      87

binom.test(table(sexo))

##
## Exact binomial test
##
## data:  table(sexo)
## number of successes = 113, number of trials = 200, p-value =
## 0.07684
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4933 0.6348
## sample estimates:
## probability of success
##                0.565
##

```

El p-valor del contraste (0.076) indica que la proporción de hembras (en el contraste por defecto, cuando se utiliza `table()` la primera categoría de la variable es la que se considera como *éxito* (*success*); si no se ha especificado ordenación, prevalece el orden alfabético; de ahí que el contraste que se ha hecho en este caso es $H_0 : \pi_H = 0,5$, siendo π_H la proporción de hembras en la población) no difiere significativamente de 0.5, y por tanto no hay diferencia significativa en las proporciones de machos y hembras. Vemos de esta forma que aunque las hembras abundan ligeramente más que los machos (56.5% frente a 43.5%), estos valores están dentro de las variación que cabe esperar por azar si ambas proporciones fueran del 50% en la población.

7. Estima la proporción de peces parasitados. Construye intervalos de confianza al 95% para dicha proporción utilizando los métodos de Wilson, Agresti-Coull y Clopper-Pearson (páginas 26-32). ¿Cuál de los tres intervalos consideras más apropiado en este caso y por qué?

Solución:

Para estimar la proporción de peces parasitados utilizamos `table()` y `prop.table()` :

```

table(ptdo)

## ptdo
##  0  1
## 170 30

prop.table(table(ptdo))

## ptdo
##  0  1
## 0.85 0.15

```

Así pues, los peces parasitados representan un 15% del total de la muestra. Para calcular los intervalos de confianza pedidos utilizamos la función `binom.confint()` de la librería `binom`. El intervalo de Clopper-Pearson corresponde al método exacto:

```

library(binom)

## Loading required package: lattice

binom.confint(30, 200, method = "wilson")

##      method x    n mean  lower  upper
## 1 wilson 30 200 0.15 0.1071 0.2061

binom.confint(30, 200, method = "agresti")

##           method x    n mean  lower  upper
## 1 agresti-coull 30 200 0.15 0.1067 0.2065

binom.confint(30, 200, method = "exact")

##      method x    n mean  lower  upper
## 1  exact 30 200 0.15 0.1035 0.2072

```

Los tres intervalos son similares, si bien el tamaño de la muestra ($n = 200$) y el hecho de que haya más de 5 observaciones en cada grupo (parasitados y no parasitados), aconsejan utilizar el intervalo de Wilson, que es ligeramente más estrecho que los otros.

8. Estima la proporción de peces parasitados respectivamente entre las hembras y entre los machos. ¿Qué sexo está proporcionalmente más parasitado en esta muestra? (usar las funciones `table` y `prop.table`. Si no recuerdas la sintaxis, utiliza la ayuda de R). Construye e interpreta sendos intervalos de confianza (al 95%) para la diferencia y el cociente entre ambas proporciones (pags. 32-35).

NOTA: en el ejemplo que se muestra en la guía se emplea la función `prop.test(c(30, 28), c(160, 125))` donde los valores numéricos `c(30, 28)` y `c(160, 125)` corresponden, respectivamente, al número de éxitos y al número de observaciones en cada grupo. Si, como ocurre con los datos de esta práctica, se especifica una variable (`ptdo`) que especifica éxito/fracaso y otra (`sexo`) que especifica el grupo, Rhace directamente el recuento de éxitos y tamaños de grupo utilizando `prop.test(table(ptdo, sexo))`

Solución:

Estimamos la proporción de peces parasitados por sexo:

```

parasitado = factor(ptdo, levels = c(1, 0), labels = c("Sí",
  "No"), ordered = TRUE)
table(sexo, parasitado)

##           parasitado
## sexo      Sí      No
## Hembra  12    101
## Macho   18     69

prop.table(table(sexo, parasitado), 1)

##           parasitado
## sexo      Sí      No
## Hembra 0.1062 0.8938
## Macho  0.2069 0.7931

```

La primera tabla nos muestra el número de machos y hembras parasitados. La segunda tabla calcula las proporciones por filas y por tanto nos da la proporción de parasitados y no parasitados en hembras (primera fila) y la proporción de parasitados y no parasitados en machos (segunda fila). Vemos así que, en términos relativos, la proporción de parasitados entre los machos (20.69 %) prácticamente duplica a dicha proporción entre las hembras (10.62 %). Para determinar si esta diferencia es significativa efectuamos el test de la chi-cuadrado:

```
prop.test(table(sexo, parasitado))

##
## 2-sample test for equality of proportions with continuity
## correction
##
## data:  table(sexo, parasitado)
## X-squared = 3.16, df = 1, p-value = 0.07548
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.2132  0.0118
## sample estimates:
## prop 1 prop 2
## 0.1062 0.2069
##
```

El p-valor del test (0.07548) indica que éste no ha resultado significativo a un nivel α del 5%, por lo que aceptamos que no existen diferencias en la proporción de peces parasitados entre machos y hembras. El comando `prop.test()` nos ha calculado también un intervalo de confianza al 95 % para la diferencia entre π_H y π_M (proporciones de parasitados, respectivamente, en machos y hembras):

$$\pi_H - \pi_M \in [-0,2132, 0,0118]$$

Para dar un intervalo de confianza para el cociente de estas proporciones utilizamos la función `riskcoreci()` de la librería `PropCIs` . Hemos de tener en cuenta que para utilizar esta función hay que consignar el total de sujetos de cada sexo:

```
library(PropCIs)
table(sexo)

## sexo
## Hembra Macho
##    113    87

riskcoreci(12, 113, 18, 87, conf = 0.95)

##
##
##
## data:
##
## 95 percent confidence interval:
## 0.2639 0.9944
##
```

Así pues, el intervalo al 95 % para el cociente de proporciones es:

$$\frac{\pi_H}{\pi_M} \in [0,2639, 0,9944]$$

Obsérvese que este intervalo *no contiene al 1*; como está estrictamente por debajo de 1, nos lleva a concluir que un un 95 % de confianza la proporción de parasitados entre las hembras es inferior que entre los machos; ello entra en contradicción con el intervalo anterior para la diferencia, que contiene al 0 y por tanto lleva a la conclusión de que ambas proporciones no muestran diferencias significativas (tal como nos indica además el p-valor del test que hemos realizado). ¿A qué se debe esta contradicción entre resultados?. Fundamentalmente a que ambos intervalos son *asintóticos* (esto es, son válidos sólo para muestras grandes). Aunque puede argumentarse que las muestras de machos y hembras (87 y 113 ejemplares respectivamente) son grandes, es obvio que no lo son lo suficiente para que ambos procedimientos converjan a la misma solución: el intervalo para el cociente se queda por debajo de 1, pero por muy poco; y el intervalo para la diferencia contiene al cero, pero también por muy poco. Una situación como ésta indica que la información con la que contamos es insuficiente. El principio de prudencia aconseja aceptar provisionalmente que ambas proporciones no muestran diferencias significativas, y que es aconsejable adquirir más información (tomar más datos).

9. Suponiendo que el número de larvas de anisakis en cada pez sigue una distribución de Poisson, estima su parámetro y acompaña la estimación de un intervalo de confianza al 90 % (págs 36 y 37).

Solución:

El número de larvas encontrado en los distintos peces se distribuye de acuerdo a la siguiente tabla de frecuencias:

```
table(larvas)
```

```
## larvas
##   0   3   4   5   6   7   8   9
## 170  4  2  4  2  3  9  6
```

siendo el total de larvas:

```
sum(larvas)
```

```
## [1] 199
```

El parámetro λ de la distribución de Poisson corresponde a la esperanza de dicha variable. Por tanto, podemos estimarlo mediante la media muestral:

```
mean(larvas)
```

```
## [1] 0.995
```

y el intervalo de confianza es de la forma:

```
n = 200
T = 199
c(qchisq(0.025, 2 * T), qchisq(0.975, 2 * (T + 1)))/(2 * n)
## [1] 0.8616 1.1433
```

10. Construye una tabla con el número de animales parasitados según sexo, mediante `table(sexo, ptdo)`. Construye e interpreta la tabla en frecuencias relativas `prop.table(table(sexo, ptdo), 1)`. Utilizando el comando `prop.test`, ¿muestran los datos evidencia significativa al 5% de que la proporción de peces parasitados es inferior en las hembras?

Solución:

La tabla y su interpretación ya se han llevado a cabo en el problema 8. Siendo π_H la proporción de parasitados entre las hembras (que se estima en un 10.62% tal como hemos visto en el problema 8) y π_M la proporción de parasitados entre los machos (estimada en un 20.69%), la pregunta que se plantea ahora (¿existe evidencia suficiente de que $\pi_H < \pi_M$?) puede expresarse mediante el contraste:

$$\begin{cases} H_0 : \pi_H \geq \pi_M \\ H_1 : \pi_H < \pi_M \end{cases}$$

A diferencia de lo que hicimos en la pregunta 8, ahora planteamos la alternativa $\pi_H < \pi_M$:

```
##
## 2-sample test for equality of proportions with continuity
## correction
##
## data: table(sexo, parasitado)
## X-squared = 3.16, df = 1, p-value = 0.03774
## alternative hypothesis: less
## 95 percent confidence interval:
## -1.000000 -0.004649
## sample estimates:
## prop 1 prop 2
## 0.1062 0.2069
##
```

Obsérvese que el p-valor (0.0377) indica que, efectivamente, existe evidencia suficiente para rechazar H_0 y concluir que las hembras son menos parasitadas que los machos.

Estamos aquí en una aparente contradicción con el resultado de la pregunta 8. Con los mismos datos, allí concluíamos que no había evidencia suficiente de que la proporción de parasitados fuese diferente entre sexos, mientras que aquí concluimos que hay evidencia suficiente de que las hembras son menos parasitadas que los machos. ¿En cuál de los dos problemas hemos tomado la decisión correcta y en cuál nos hemos equivocado? Por extraño que resulte, no nos hemos equivocado en ninguno. Las respuestas son distintas *porque las preguntas eran distintas*:

- En la pregunta 8 se pedía decidir si había evidencia suficiente de que las proporciones fuesen *distintas*. Pero claro, ser *distintas* implica que puede ser que $\pi_H < \pi_M$ o que $\pi_H > \pi_M$, y el contraste está construido de tal forma que sea sensible a ambas posibilidades, lo que a su vez implica que es menos sensible a cada una de ellas por separado.
- En esta pregunta sólo se desea decidir si hay evidencia suficiente de que la proporción de parasitados es *menor en las hembras*; no hay que contemplar como posible resultado tener que rechazar H_0 si $\pi_H > \pi_M$; por tanto la evidencia se acumula sólo en un sentido.

11. Crear la variable `oriental` del siguiente modo:

```
oriental=ifelse(isla=="FV"|isla=="LZ"|isla=="GC", TRUE, FALSE)
```

(esto es, hemos construido una variable que vale `TRUE` si la isla pertenece a la zona oriental del archipiélago, y `FALSE` en caso contrario). ¿Puede aceptarse, con un 5% de significación, que en

las islas de la parte occidental del archipiélago hay la misma proporción de machos y hembras?
¿Difiere de manera significativa la proporción de hembras entre las dos zonas?

Solución:

En primer lugar definimos la variable *oriental* y evaluamos la proporción de machos y hembras en cada zona:

```
oriental = ifelse(isla == "FV" | isla == "LZ" | isla == "GC",
  TRUE, FALSE)
table(oriental, sexo)

##           sexo
## oriental Hembra Macho
##   FALSE      47     41
##   TRUE       66     46

prop.table(table(oriental, sexo), 1)

##           sexo
## oriental Hembra Macho
##   FALSE 0.5341 0.4659
##   TRUE  0.5893 0.4107
```

Vemos que la distribución por sexos es muy similar en ambas regiones, si bien en la muestra procedente de la región oriental las hembras son ligeramente más abundantes en comparación con la región occidental (58.93 % frente a 53.41 %). Para contrastar si esta diferencia es significativa, realizamos el test de la chi-cuadrado:

```
prop.test(table(oriental, sexo))

##
## 2-sample test for equality of proportions with continuity
## correction
##
## data:  table(oriental, sexo)
## X-squared = 0.4069, df = 1, p-value = 0.5235
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.20377  0.09338
## sample estimates:
## prop 1 prop 2
## 0.5341 0.5893
##
```

El p-valor del test (0.5235) indica que la diferencia entre regiones no es significativa, y que por tanto se puede asumir que la proporción de hembras es idéntica en ambas regiones.

Por último, para decidir si en la parte occidental del archipiélago hay la misma proporción de machos que de hembras, habremos de contrastar si puede aceptarse que en dicha región la proporción de hembras es el 50 % (en cuyo caso la proporción de machos será el otro 50 %, y no hay diferencias en la proporción de individuos de cada sexo):


```
prop.test(table(sexo[oriental == F]))

##
## 1-sample proportions test with continuity correction
##
## data:  table(sexo[oriental == F]), null probability 0.5
## X-squared = 0.2841, df = 1, p-value = 0.594
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4251 0.6401
## sample estimates:
##      p
## 0.5341
##
```

El p-valor del test (0.594) muestra que la proporción de hembras no difiere significativamente del 50%. Por tanto, puede aceptarse la hipótesis de que en dicha región ambos sexos se reparten a partes iguales.

12. Los modelos pesqueros tradicionales indican que al menos la cuarta parte de los sargos de archipiélago se encuentra en aguas de Gran Canaria. Suponiendo que el muestreo se ha realizado de manera completamente aleatoria, ¿permiten los datos desmentir esta hipótesis?

Solución:

Si llamamos π_{GC} a la proporción de sargos que habitan en aguas de Gran Canaria, hemos de resolver el contraste:

$$\begin{cases} H_0 : \pi_{GC} = 0,25 \\ H_1 : \pi_{GC} \neq 0,25 \end{cases}$$

La proporción de sargos estimada en cada una de las islas es:

```
table(isla)

## isla
## FV GC HI LG LP LZ TF
## 32 48 15  9 24 32 40

prop.table(table(isla))

## isla
##      FV      GC      HI      LG      LP      LZ      TF
## 0.160 0.240 0.075 0.045 0.120 0.160 0.200
```

Como vemos, en Gran Canaria se ha capturado un 24% de los sargos de la muestra (48 sargos de un total de 200). Para realizar el contraste pedido podemos utilizar la función `binom.test()`, especificando como hipótesis nula que la proporción de éxitos es $p = 0,25$:

```

binom.test(48, 200, p = 0.25)

##
## Exact binomial test
##
## data: 48 and 200
## number of successes = 48, number of trials = 200, p-value =
## 0.8066
## alternative hypothesis: true probability of success is not equal to 0.25
## 95 percent confidence interval:
## 0.1826 0.3053
## sample estimates:
## probability of success
## 0.24
##

```

El p-valor del test (0.8066) indica que no existe evidencia suficiente para rechazar la hipótesis nula. Por tanto los datos observados son perfectamente compatibles con que el Gran Canaria se realiza el 25 % de las capturas de esta especie.

Otra manera de realizar el contraste es mediante el *test chi-cuadrado de bondad de ajuste*, que compara una distribución observada de datos con la distribución teórica bajo la hipótesis nula. En este caso, la distribución observada es `c(48, 152)` (48 individuos en Gran Canaria y 152 en el resto del archipiélago); la distribución teórica puede expresarse de dos maneras:

- en términos porcentuales: `c(0.25, 0.75)` (se espera un 25 % de sujetos en Gran Canaria y un 75 % en el resto del archipiélago).
- en términos absolutos: `c(50, 150)` (si la hipótesis nula especifica un 25 % en Gran Canaria y un 75 % en el resto del archipiélago, de un total de 200 peces cabría esperar, bajo dicha hipótesis que el 25 %, esto es, 50, procedan de Gran Canaria y un 75 %, esto es, 150, del resto del archipiélago). En este caso hay que especificar la opción `rescale.p=T`.

De esta forma, el contraste puede realizarse de las dos siguientes maneras, ambas equivalentes tal como puede comprobarse ya que producen el mismo p-valor:

```

chisq.test(x = c(48, 152), p = c(0.25, 0.75))

##
## Chi-squared test for given probabilities
##
## data: c(48, 152)
## X-squared = 0.1067, df = 1, p-value = 0.744
##

chisq.test(x = c(48, 152), p = c(50, 150), rescale.p = T)

##
## Chi-squared test for given probabilities
##
## data: c(48, 152)
## X-squared = 0.1067, df = 1, p-value = 0.744
##

```

El p-valor 0.744 indica que puede aceptarse H_0 , esto es, que el 25 % de las capturas se produce en Gran Canaria.

Como vemos, el test chi-cuadrado y el test binomial han producido el mismo resultado, aceptar H_0 . La diferencia entre ambos obedece a que el primero es exacto y el segundo aproximado (la aproximación mejora cuanto más grande sea la muestra). En este caso, en el que sólo hay dos categorías (el pez procede de Gran Canaria, SÍ o NO) es recomendable utilizar el test exacto. Sin embargo si hubiese más categorías (por ejemplo, que la hipótesis nula especificase que el 25 % procede de Gran Canaria, el 15 % de Tenerife y el 60 % restante del resto del archipiélago) sólo podría utilizarse el test chi-cuadrado. En el ejemplo citado, el contraste se resolvería mediante (véase en la tabla de frecuencias anterior que a Tenerife corresponden 40 peces):

```
chisq.test(x = c(48, 40, 112), p = c(0.25, 0.15, 0.6))

##
## Chi-squared test for given probabilities
##
## data:  c(48, 40, 112)
## X-squared = 3.947, df = 2, p-value = 0.139
##
```

y el p-valor (0.139) indicaría que puede aceptarse la hipótesis nula. Pueden verse más ejemplos del test de bondad de ajuste chi-cuadrado, así como una pequeña justificación del mismo, en <http://ljk.imag.fr/membres/Bernard.Ycart/emel/cours/ts/node8.html>, o también en http://www.ccee.edu.uy/ensenian/catest2/Material/PRACTICA_13_SOL09.pdf.

13. Utilizando el comando `power.t.test` determinar el tamaño de muestra necesario si se desea detectar una diferencia de 3 gramos en el peso de las gónadas entre machos y hembras con una significación del 5% y una potencia del 80%. (**Nota:** potencia = $1 - P(\text{error II})$). Utilizar como estimación piloto de la desviación típica el valor estimado en la muestra que hemos utilizado en esta práctica.

Solución:

La sintaxis del comando `power.t.test()` es la siguiente:

```
power.t.test(power = 0.8, delta = 3, sig.level = 0.05, sd = sd(pgon,
  na.rm = T))

##
## Two-sample t test power calculation
##
##          n = 744
##        delta = 3
##         sd = 20.64
##    sig.level = 0.05
##     power = 0.8
## alternative = two.sided
##
## NOTE: n is number in *each* group
##
```

Por tanto, para las condiciones especificadas se necesitaría un tamaño de muestra de 744 sujetos de cada sexo.

Esta función puede utilizarse también para calcular la potencia de un test para detectar una diferencia *delta* con un tamaño de muestra especificado *n*. Por ejemplo, si el tamaño de muestra fuese de $n = 200$, la potencia para detectar una diferencia de 3 gramos con una significación del 5% se obtendría mediante:

```
power.t.test(n = 200, delta = 3, sig.level = 0.05, sd = sd(pgon,
  na.rm = T))

##
##      Two-sample t test power calculation
##
##              n = 200
##             delta = 3
##             sd = 20.64
##      sig.level = 0.05
##             power = 0.305
## alternative = two.sided
##
## NOTE: n is number in *each* group
##
```

Señalemos, por último, que como la variable *pgon* no es normal, los valores de tamaño muestral y potencia calculados de esta forma no son válidos (ya que el t-test requiere la normalidad de los datos) y deben tomarse como meramente orientativos en el diseño de un muestreo.

14. Construye un histograma del peso de las gónadas para cada sexo. ¿Crees que el peso de las gónadas sigue una distribución normal?. Transforma el peso de las gónadas a escala logarítmica y traza de nuevo un histograma ¿Se ha conseguido que los datos se aproximen a la normalidad? Se muestran a continuación los comandos para esta tarea:

```
par(mfrow = c(2, 2))

hist(pgon[sexo == "Macho"])

hist(pgon[sexo == "Hembra"])

hist(log(pgon[sexo == "Macho"]))

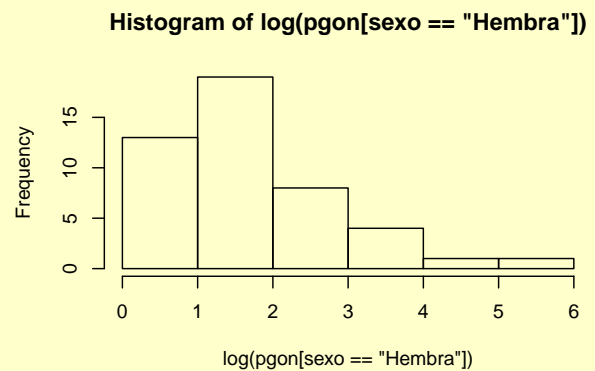
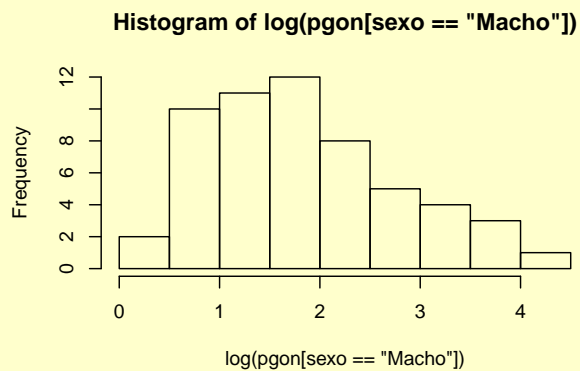
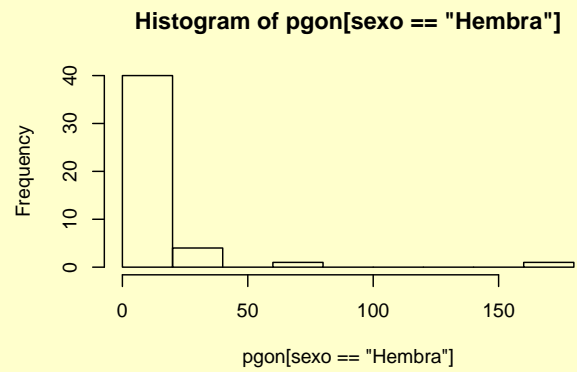
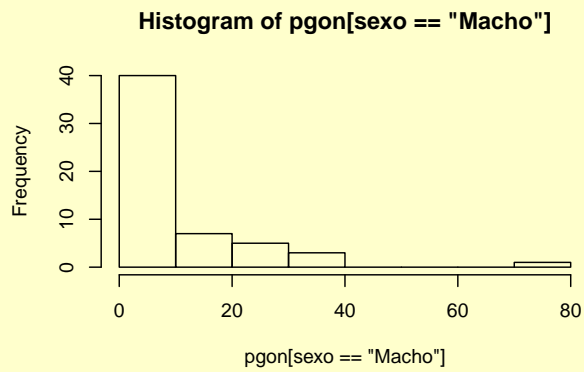
hist(log(pgon[sexo == "Hembra"]))
```

En la pregunta 1, apartado c, se ha utilizado el *t.test* para comparar el peso de las gónadas entre machos y hembras. ¿Consideras que es correcto?. Repite dicho apartado utilizando el test de Wilcoxon (mediante el comando *wilcox.test*).

Solución:

Dibujamos los histogramas pedidos:

```
par(mfrow = c(2, 2))
hist(pgon[sexo == "Macho"])
hist(pgon[sexo == "Hembra"])
hist(log(pgon[sexo == "Macho"]))
hist(log(pgon[sexo == "Hembra"]))
```



En la resolución del apartado 1c ya comprobamos, utilizando el test de Shapiro-Wilk, que el peso de las gónadas no sigue una distribución normal en ninguno de los dos sexos, por lo que resultaba incorrecto el uso del *t.test* y, de hecho, ya allí empleamos el test de Wilcoxon. Los histogramas para los datos transformados a logaritmo sigue mostrando cierta asimetría por lo que cabe esperar que tampoco sigan distribución normal. Empleando el test de Shapiro-Wilk:

```
shapiro.test(log(pgon[sexo == "Macho"]))

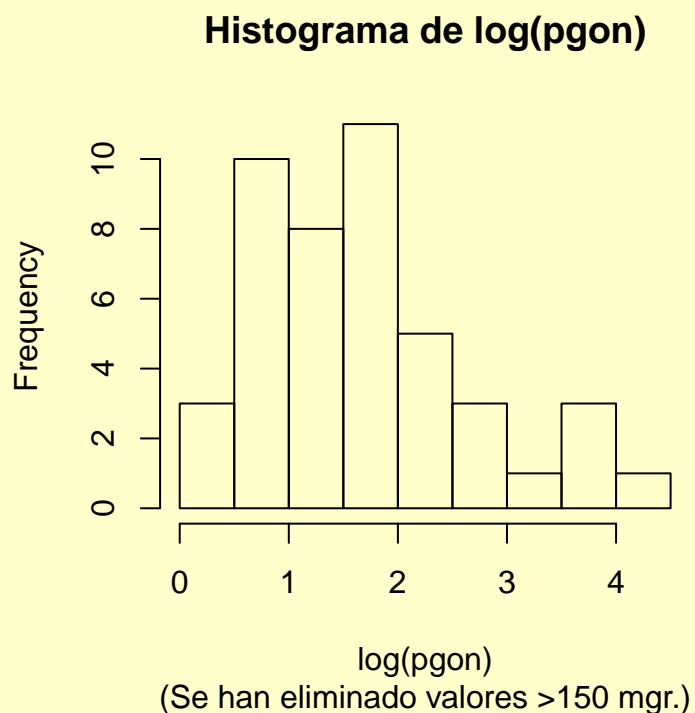
##
## Shapiro-Wilk normality test
##
## data: log(pgon[sexo == "Macho"])
## W = 0.9681, p-value = 0.1445
##
```

```
shapiro.test(log(pgon[sexo == "Hembra"]))

##
## Shapiro-Wilk normality test
##
## data: log(pgon[sexo == "Hembra"])
## W = 0.9055, p-value = 0.001241
##
```

vemos que en los machos puede aceptarse la normalidad (p -valor=0.1445) pero no así en las hembras (0.0012). Si revisamos el *boxplot* que se hizo en la pregunta 1c, vemos que hay un valor de una hembra que está muy alejado del resto; dicho valor puede deberse a múltiples circunstancias: desde un mero error de transcripción (el valor debía ser 17.12 y se transcribió como 171.2), un valor mal medido, esa hembra que presentaba una anomalía que la hacía diferente al resto, ... Si el investigador tiene razones para pensar que ese valor se ha "infiltrado" en la muestra pero que realmente no corresponde a la variable que se está midiendo, puede prescindir de él. Si en este caso eliminamos ese valor anómalo, el histograma que queda para las hembras es de la forma:

```
hist(log(pgon[sexo == "Hembra" & pgon < 150]), main = "Histograma de
log(pgon)",
      xlab = "log(pgon)", sub = "(Se han eliminado valores >150 mgr.)")
```



y el test de Shapiro-Wilk produce el resultado siguiente:

```
shapiro.test(log(pgon[sexo == "Hembra" & pgon < 150]))

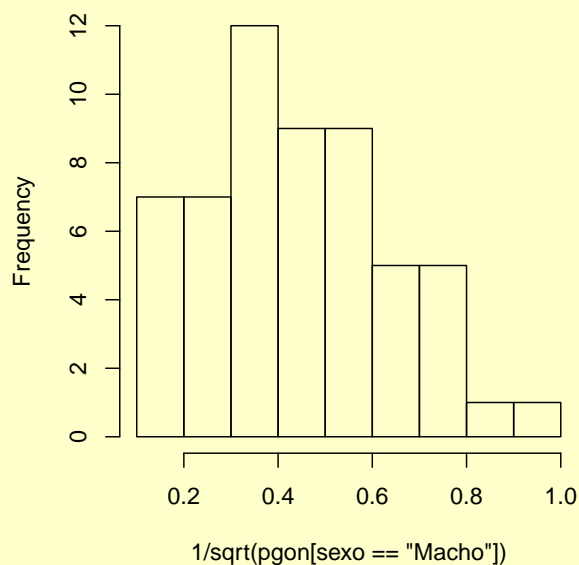
##
## Shapiro-Wilk normality test
##
## data: log(pgon[sexo == "Hembra" & pgon < 150])
## W = 0.9268, p-value = 0.007244
##
```

El p-valor (0.007) sigue siendo demasiado pequeño para aceptar la normalidad de la variable transformada.

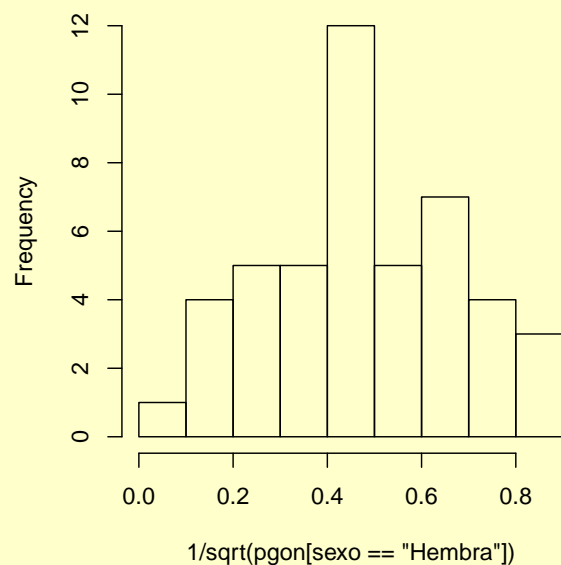
En la práctica se suelen usar las funciones $\log(x)$, x^2 , \sqrt{x} , $1/x$, $1/\sqrt{x}$, $1/x^2$ para tratar de convertir en normales variables que inicialmente no lo son. Estas funciones forman parte de una clase más amplia, conocida como *transformaciones de Box y Cox* que tienen como objetivo, precisamente, conseguir que los datos se conviertan en normales, si bien este objetivo no siempre es posible conseguirlo. La elección de la transformación óptima para conseguir normalidad se encuentra implementada en la librería *MASS* si bien su uso queda fuera de este curso. A modo de ejemplo, podemos comprobar que la variable *pgon* puede transformarse en normal mediante el uso de la función $1/\sqrt{x}$:

```
par(mfrow = c(1, 2))
hist(1/sqrt(pgon[sexo == "Macho"]))
hist(1/sqrt(pgon[sexo == "Hembra"]))
```

Histogram of 1/sqrt(pgon[sexo == "Macho"])



Histogram of 1/sqrt(pgon[sexo == "Hembra"])



```
shapiro.test(1/sqrt(pgon[sexo == "Macho"]))
```

```
##
## Shapiro-Wilk normality test
##
## data: 1/sqrt(pgon[sexo == "Macho"])
## W = 0.9727, p-value = 0.233
##
```

```
shapiro.test(1/sqrt(pgon[sexo == "Hembra"]))
```

```
##
## Shapiro-Wilk normality test
##
## data: 1/sqrt(pgon[sexo == "Hembra"])
## W = 0.9813, p-value = 0.6605
##
```

Vemos, pues, que con los datos así transformados obtenemos variables normales. Si queremos evaluar si *pgon* muestra la misma tendencia central en machos y hembras podríamos llevar a cabo un **t.test()** no sobre la variable original *pgon*, sino sobre la variable transformada:

```
t.test(1/sqrt(pgon) ~ sexo)

##
## Welch Two Sample t-test
##
## data: 1/sqrt(pgon) by sexo
## t = 0.8678, df = 94.31, p-value = 0.3877
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.04406 0.11248
## sample estimates:
## mean in group Hembra mean in group Macho
## 0.4754 0.4412
##
```

El p-valor (0.3844) implica que no se detectan diferencias significativas en los valores medios de esta variable entre sexos, de donde podríamos deducir que no hay diferencias significativas entre sexos en la tendencia central de la variable original *pgon*. En cualquier caso hay que ser cautos a la hora de interpretar este tipo de resultados, ya que los promedios de los valores transformados (en este caso las medias de las inversas de las raíces cuadradas) pueden tener muy poco que ver con las medias de los valores originales.