Capítulo 5

Inferencia Estadística II: Estimación por Intervalos de Confianza

Problemas

1. Las larvas de algunas mariposas monarchi concentran glucósidos a partir de plantas de algoncillo. En cierta localidad se han capturado 14 de estas larvas y se les han medido las concentraciones de glucósidos, resultando una concentración media de 0.2 con una desviación típica de 0.012. Suponiendo que esta variable es normal:

   a) Construir sendos intervalos de confianza al 95 % para las verdaderas media y varianza de la población.

   b) A partir del intervalo para la media, ¿es verosímil que la verdadera media de la población pueda ser 0.15?.

   c) El mismo experimento se ha realizado con 12 larvas de mariposas de otra especie, obteniéndose una media de 0.28 con desviación típica de 0.02. Construir un intervalo de confianza al 95 % para la diferencia media de las concentraciones entre ambas especies. ¿podría aceptarse que la concentración media poblacional de glucósidos en la segunda especie es igual a la de la primera especie?.

> **Solución:**

a) El intervalo de confianza a nivel $1-\alpha$ para la media de una variable con distribución normal $N(\mu, \sigma)$ es de la forma:

$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

Asimismo, para la varianza el intervalo es de la forma:

$$\sigma^2 \in \left[\frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

En este caso: $\overline{X} = 0{,}2$, $S = 0{,}012$, $n = 14$. Para $\alpha = 0{,}05$ obtenemos (consultando las tablas) $t_{n-1,\alpha/2} = t_{13,0{,}025} = 2{,}16$, $\chi^2_{n-1,\alpha/2} = \chi^2_{13,0{,}025} = 24{,}736$ y $\chi^2_{n-1,1-\alpha/2} = \chi^2_{13,0{,}975} = 5{,}009$. También podemos obtener los valores directamente en R mediante:

$$t_{n-1,\alpha/2} = t_{13,0{,}025} = \texttt{qt(0.975,13)} = 2{,}1604$$

$$\chi^2_{n-1,\alpha/2} = \chi^2_{13,0{,}025} = \texttt{qchisq(0.975,13)} = 24{,}736$$

$$\chi^2_{n-1,1-\alpha/2} = \chi^2_{13,0{,}975} = \texttt{qt(0.025,13)} = 5{,}0088$$

Sustituyendo estos valores en las expresiones de los intervalos anteriores obtenemos:

$$\mu \in [0{,}19307, 0{,}20693]$$

$$\sigma^2 \in [0{,}000076, 0{,}000333]$$

Podemos hallar un intervalo para la desviación típica mediante la raíz cuadrada de los extremos del intervalo para la varianza:

$$\sigma \in \left[\sqrt{0{,}000076}, \sqrt{0{,}000333}\right] = [0{,}008699, 0{,}018237]$$

Estos intervalos podían haberse obtenido fácilmente en R mediante:

```
> media = 0.2
> s = 0.012
> n = 14
> a2 = 1 - 0.05/2
> media + c(-1, 1) * qt(a2, n - 1) * s/sqrt(n)
[1] 0.19307 0.20693
> (n - 1) * s^2/c(qchisq(a2, n - 1), qchisq(1 - a2, n))
[1] 0.00007568 0.00033258
```

```
> sqrt((n - 1) * s^2/c(qchisq(a2, n - 1), qchisq(1 - a2,
    n)))
[1] 0.0086994 0.0182368
```

b) A la vista del intervalo de confianza para la media, podemos estar muy seguros (al 95 % de confianza) de que *el verdadero valor de la media de esta población* está comprendido en el intervalo $[0{,}19307, 0{,}20693]$. Como el valor 0.15 no cae en este intervalo, podemos concluir que no resulta verosímil que la verdadera media de la población pueda ser 0.15.

c) Podemos asumir razonablemente que ambas muestras son independientes. Asimismo, no tenemos razones para pensar que las varianzas en ambas poblaciones puedan ser iguales, por lo que calcularemos el intervalo de confianza para la diferencia de medias entre dos poblaciones normales con muestras independientes:

$$\mu_1 - \mu_2 \in \left[(\bar{X}_1 - \bar{X}_2) \pm t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

donde:

$$n = REDONDEO \left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2} \right)^2 \frac{1}{n_2 - 1}} \right]$$

Los datos disponibles son $\overline{X}_1 = 0{,}2$, $\overline{X}_2 = 0{,}28$, $S_1 = 0{,}012$, $S_2 = 0{,}02$, $n_1 = 14$, $n_2 = 12$. Sustituyendo en las expresiones anteriores obtenemos que $n = 17$. Buscamos en la tabla $t_{17,0{,}025} = 2{,}1098$ y sustituímos en la expresión del intervalo de confianza, obteniendo finalmente:

$$\mu_1 - \mu_2 \in [-0{,}093934, -0{,}066066]$$

Como el intervalo es íntegramente negativo, podemos estar bastante seguros (con una confianza del 95 %) de que la diferencia $\mu_1 - \mu_2$ es negativa, lo que significa que $\mu_1 < \mu_2$ (concretamente, μ_1 es entre 0.066066 y 0.093934 unidades menor que μ_2). Por tanto, no podemos aceptar la suposición de que ambas especies tienen la misma concentración media de glucósidos.

2. Con objeto de determinar cual de dos diferentes técnicas de cría de peces produce mayor rendimiento, se ha medido la producción, en toneladas, durante 12 periodos de cultivo para la técnica A y durante 10 periodos para la B, con los siguientes resultados:

Técnica A	8.93	9.54	10.32	6.99	8.56	8.67	9.72	7.76	8.95	9.32	8.59	9.78
Técnica B	5.69	7.56	6.88	10.26	9.57	7.88	8.95	9.35	6.58	7.32		

Estimar mediante un intervalo de confianza al 95 % la diferencia media en rendimiento entre ambas técnicas. ¿Apuntan estos datos a que alguna técnica produce mejor rendimiento que otra? Justificar la respuesta.

Solución:

Podemos suponer que los datos en ambos grupos son independientes, por lo que para estimar $\mu_A - \mu_B$ utilizaremos el mismo procedimiento que en el apartado (c) del problema anterior. Calculamos en primer lugar las medias y desviaciones típicas en cada grupo:

```
> A = c(8.93, 9.54, 10.32, 6.99, 8.56, 8.67, 9.72, 7.76,
     8.95, 9.32, 8.59, 9.78)
> B = c(5.69, 7.56, 6.88, 10.26, 9.57, 7.88, 8.95, 9.35,
     6.58, 7.32)
> mediaA = mean(A)
> sA = sd(A)
> mediaB = mean(B)
> sB = sd(B)
> mediaA

[1] 8.9275

> mediaB

[1] 8.004

> sA

[1] 0.9173

> sB

[1] 1.4748
```

Calculamos ahora los grados de libertad de la t de Student:

$$n = REDONDEO\left[\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\frac{1}{n_1-1}+\left(\frac{s_2^2}{n_2}\right)^2\frac{1}{n_2-1}}\right] = 15$$

Buscamos en la tabla $t_{15,0,025} = 2{,}1314$ y sustituímos en la expresión del intervalo de confianza, obteniendo finalmente:

$$\mu_1 - \mu_2 \in \left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n,\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right] = [-0{,}21959, 2{,}0666]$$

Como el intervalo de confianza contiene al 0 (tiene una parte positiva y una negativa), ello significa que con la información disponible cabe la posibilidad de que la diferencia $\mu_A - \mu_B$ sea positiva, negativa o nula; en otras palabras, la muestra no contiene información suficiente para diferenciar claramente entre los rendimientos de ambos tipos de cultivo. Por tanto no hemos detectado diferencias claras entre ambos rendimientos, así que la decisión más prudente es aceptar que ninguna de las técnicas de cultivo ha demostrado ser superior a la otra.

El intervalo de confianza puede obtenerse de manera muy sencilla en R mediante la sintaxis:

```
> t.test(A, B)

        Welch Two Sample t-test

data:  A and B
t = 1.722, df = 14.505, p-value = 0.1063
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.223  2.070
sample estimates:
mean of x mean of y
   8.9275    8.0040
```

Como vemos, el intervalo calculado por R difiere ligeramente del que hemos calculado antes; ello se debe fundamentalmente a que R no redondea el valor de n pues puede calcular exactamente el valor de $t_{n,\alpha}$ incluso para valores de n no enteros. Por ello, el intervalo calculado por R es más preciso (y preferible al que hemos obtenido haciendo las cuentas "a mano"). En este caso, utilizar el intervalo proporcionado por R no cambia para nada la interpretación que hemos hecho (sigue conteniendo al cero).

3. La siguiente tabla muestra los resultados de un experimento para comparar dos métodos diferentes de medición del contenido en KCO_3 (en $\mu gr/gr$) de las algas de cierta especie. Para ello se eligieron veinte muestras de algas, y se separó cada una en dos fracciones, utilizándose un método de medida de contenido en KCO_3 en cada fracción. Obtener intervalos de confianza al 95 % para el contenido medio calculado con cada método, así como para la diferencia entre ambos. ¿Las diferencias entre las medidas obtenidas por ambos métodos pueden achacarse simplemente al azar o muestran los datos evidencia de que un método tiende a dar valores mayores que el otro?

Método A	Método B
58.61	63.32
98.72	105.87
77.28	75.89
73.52	74.5
93.12	95.09
75.79	83.76
66.58	71.06
66	66.18
83.02	84.33
108.49	119.54
35.92	29.85
109.44	123.59

Solución:

Para calcular un intervalo de confianza para cada método hemos de calcular medias y desviaciones típicas en cada grupo:

```
> A = c(58.61, 98.72, 77.28, 73.52, 93.12, 75.79, 66.58,
     66, 83.02, 108.49, 35.92, 109.44)
> B = c(63.32, 105.87, 75.89, 74.5, 95.09, 83.76, 71.06,
     66.18, 84.33, 119.54, 29.85, 123.59)
> mean(A)

[1] 78.874

> sd(A)

[1] 21.402

> mean(B)

[1] 82.748

> sd(B)

[1] 25.962
```

y utilizar en cada caso un intervalo de la forma:

$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right]$$

En este caso $t_{11,0,025} = 2{,}201$, y sustituyendo los valores anteriores obtenemos:

$$\begin{aligned} \mu_A &\in [65{,}276, 92{,}473] \\ \mu_B &\in [66{,}253, 99{,}244] \end{aligned}$$

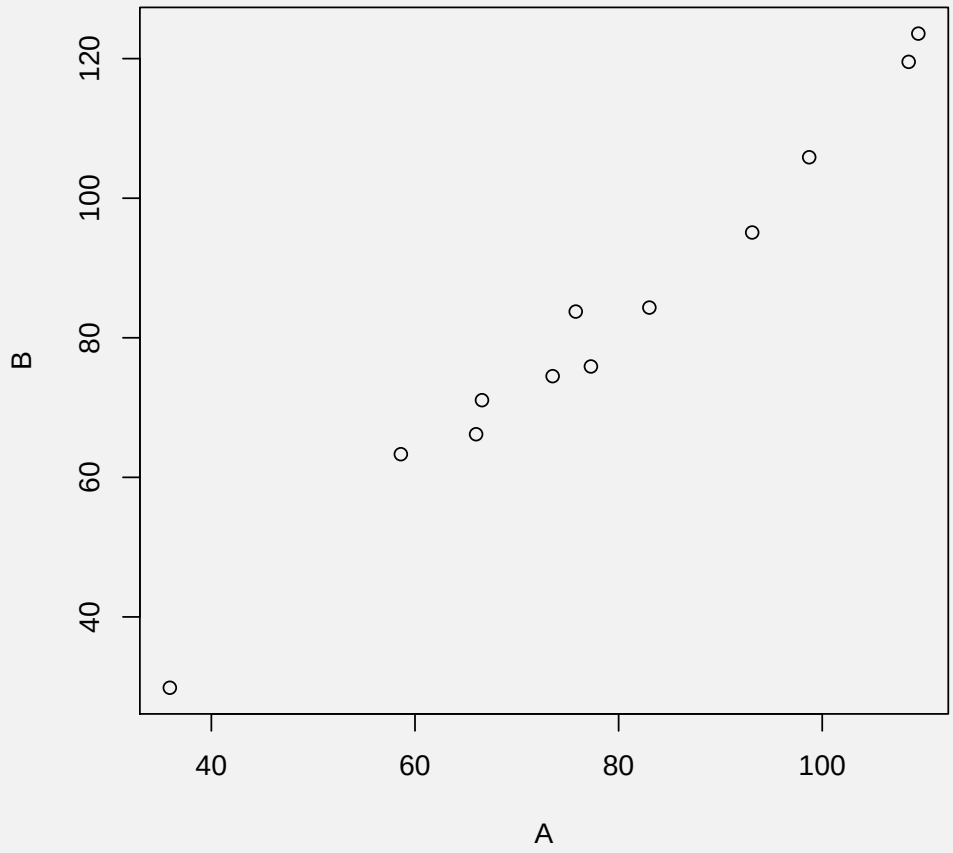
Nótese que los intervalos se solapan bastante, lo que nos lleva a pensar que efectivamente ambos métodos está midiendo esencialmente lo mismo. Estos intervalos pueden obtenerse también directamente en R mediante:

```
> t.test(A)

        One Sample t-test

data:  A
t = 12.766, df = 11, p-value = 6.142e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 65.276 92.473
sample estimates:
mean of x
   78.874

> t.test(B)

        One Sample t-test

data:  B
t = 11.041, df = 11, p-value = 2.724e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 66.253 99.244
sample estimates:
mean of x
   82.748
```

El experimento descrito corresponde a un diseño de muestras emparejadas, ya que cada una de las muestras originales se separó en dos porciones, que fueron analizadas cada una por un método

distinto. Si representamos gráficamente estos datos, podemos comprobar que existe una asociación clara entre las medidas obtenidas con ambos métodos, tal como cabía esperar ya que en cada caso se mide exactamente la misma alga; cuando el método A mide una baja cantidad de KCO_3 también lo hace el método B; y cuando A mide una cantidad alta, B también:

```
> plot(A, B)
```



Para evaluar la diferencia entre ambos métodos, podemos proceder de varias formas:

a) Calculando las diferencias:

```
> dif = A - B
> dif
 [1]  -4.71  -7.15   1.39  -0.98  -1.97  -7.97  -4.48  -0.18  -1.31
[10] -11.05   6.07 -14.15
```

```
> mean(dif)
[1] -3.8742
> sd(dif)
[1] 5.5961
```

y construyendo un intervalo de confianza para la media μ_{dif} de esta variable (nótese que como hemos definido $dif = A - B$, entonces $\mu_{dif} = \mu_A - \mu_B$):

$$\begin{aligned} \mu_{dif} &\in \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right] = \\ &= \left[-3{,}8742 - \frac{5{,}5961}{\sqrt{12}} t_{11,0{,}025}, -3{,}8742 + \frac{5{,}5961}{\sqrt{12}} t_{11,0{,}025}\right] = \\ &= [-7{,}4298, -0{,}31857] \end{aligned}$$

Como el intervalo es completamente negativo, podemos estar seguros en un 95 % de que la diferencia media es negativa y por tanto que el método A tiende en promedio a dar valores menores que el método B (aunque no podemos saber si es porque A tiende a subestimar la cantidad de KCO_3, porque B tiende a sobreestimarla, o por una combinación de ambas causas). Este intervalo podía haberse obtenido también directamente mediante:

```
> t.test(dif)

        One Sample t-test

data:  dif
t = -2.3982, df = 11, p-value = 0.03535
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -7.42976 -0.31857
sample estimates:
mean of x
  -3.8742
```

b) Constuyendo un intervalo de confianza para la diferencia de medias en muestras emparejadas. Este intervalo es de la forma:

$$\mu_1 - \mu_2 \in \left[\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{n-1,\alpha/2} \frac{S_D}{\sqrt{n}}\right]$$

siendo n el número de pares de observaciones ($n = 12$) y $S_D = \sqrt{S_1^2 + S_2^2 - 2S_{12}} = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2}$. La covarianza ($S_{12}$) es:

```
> cov(A, B)
```

```
[1] 550.37
```

y la correlación r:

```
> cor(A, B)
[1] 0.99052
```

Podemos emplear indistintamente cualquiera de las dos expresiones anteriores para calcular S_D:

$$S_D = \sqrt{S_1^2 + S_2^2 - 2S_{12}} = \sqrt{21{,}402^2 + 25{,}962^2 - 2 \cdot 550{,}37^2} = 5{,}5961$$

$$S_D = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2} = \sqrt{21{,}402^2 + 25{,}962^2 - 2 \cdot 0{,}99052 \cdot 21{,}402 \cdot 25{,}962} = 5{,}5961$$

(que, como vemos, coincide exactamente con la desviación típica calculada anteriormente para los valores de las diferencias entre A y B). Sustituyendo en la expresión del intervalo de confianza:

$$\begin{aligned} \mu_1 - \mu_2 &\in \left[(78{,}874 - 82{,}748) \pm t_{11,0{,}025}\frac{5{,}5961}{\sqrt{12}}\right] = \\ &= [-7{,}4298, -0{,}31857] \end{aligned}$$

que es exactamente el mismo intervalo que ya habíamos obtenido antes. Para obtener el intervalo para la diferencia de medias en muestras emparejadas utilizando R emplearíamos la siguiente sintaxis:

```
> t.test(A, B, paired = T)

        Paired t-test

data:  A and B
t = -2.3982, df = 11, p-value = 0.03535
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.42976 -0.31857
sample estimates:
mean of the differences
                -3.8742
```

4. Se selecciona una muestra aleatoria formada por 80 ejemplares de calamar sahariano. 27 de estos ejemplares resultaron ser machos. Estimar la proporción de machos en esta especie y dar un intervalo de confianza al 90 % para dicha proporción. En otra muestra aleatoria de 100 ejemplares de calamar de Peal se encontraron 38 machos. Estimar la diferencia entre las proporciones

de machos de ambas especies mediante un intervalo de confianza al 95 %.

Solución:

Tenemos una muestra de tamaño $n = 80$, en la que se han encontrado $n_M = 27$ machos. Por tanto, podemos estimar la proporción de machos en la población como:

$$p_M = \frac{n_M}{n} = \frac{27}{80} = 0{,}3375$$

Como el tamaño de la muestra es suficientemente grande, podemos utilizar el intervalo de Agresti-Coull:

$$\pi_M \in \left[\tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}\,(1-\tilde{\pi})}{\tilde{n}}}\right]$$

siendo (consideramos $\alpha = 0{,}10$, por lo que $z_{\alpha/2} = z_{0{,}05} = 1{,}645$):

$$\begin{aligned} \tilde{N}_E &= N_E + z^2_{\alpha/2}/2 = 27 + 1{,}645^2/2 = 28{,}353 \\ \tilde{n} &= n + z^2_{\alpha/2} = 80 + 1{,}645^2 = 82{,}706 \\ \tilde{\pi} &= \tilde{N}_E/\tilde{n} = 0{,}34282 \end{aligned}$$

Sustituyendo en la expresión del intervalo:

$$\pi_M \in \left[0{,}34282 \pm 1{,}645\sqrt{\frac{0{,}34282 \cdot (1-0{,}34282)}{82{,}706}}\right] = [0{,}25697, 0{,}42866]$$

Por tanto, estimamos que la proporción de machos en esta especie es del 33.75 %, y podemos afirmar con un 90 % de confianza que el verdadero valor de esta proporción en la población se encuentra entre el 25.697 % y el 42.866 %.

En R este intervalo puede calcularse simplemente mediante:

```
> library(binom)
> binom.confint(27, 80, method = "agresti", conf.level = 0.9)

         method  x  n   mean   lower   upper
1 agresti-coull 27 80 0.3375 0.25697 0.42866
```

Para estimar ahora la diferencia entre la proporción de machos en esta especie (población 1) y en los calamares de Peal (población 2) utilizamos el intervalo:

$$\pi_1 - \pi_2 \in \left[(\hat{\pi}_1 - \hat{\pi}_2) \pm \left(z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1\,(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2\,(1-\hat{\pi}_2)}{n_2}} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)\right]$$

Aquí $\hat{\pi}_1$ es la proporción de machos observada en la primera especie (0.3375) y $\hat{\pi}_2$ la proporción en la segunda (0.38). Los valores n_1 y n_2 son los respectivos tamaños muestrales (80 y 100). Eligiendo 95 % como nivel de significación tenemos $z_{\alpha/2} = z_{0,025} = 1{,}96$ y sustituyendo obtenemos:

$$\pi_1 - \pi_2 \in [-0{,}19442, 0{,}10942]$$

Dado que el intervalo contiene al cero (tiene una parte negativa y una positiva), concluimos que la información aportada por estas muestras *no permite asegurar que la diferencia entre las proporciones de machos en ambas especies sea significativa;* dicho de otra forma, la diferencia detectada (0.3375 frente a 0.38) puede atribuirse al azar del muestreo y no puede asegurarse que una población tenga proporcionalmente más machos que la otra.

En R este intervalo puede obtenerse directamente mediante:

```
> prop.test(c(27, 38), c(80, 100))

        2-sample test for equality of proportions with continuity
        correction

data:  c(27, 38) out of c(80, 100)
X-squared = 0.1881, df = 1, p-value = 0.6645
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.19442  0.10942
sample estimates:
prop 1 prop 2
0.3375 0.3800
```

5. En diversos estudios sobre el comportamiento territorial de la caballa se asume habitualmente que más del 60 % de las caballas se vuelven agresivas cuando se colocan otros peces en su entorno próximo.

   a) Se ha realizado un experimento con 12 caballas, observándose que en las condiciones citadas mostraron comportamiento agresivo sólo cuatro de ellas. Obtener un intervalo de confianza para la proporción poblacional de caballas agresivas a partir de estos datos. El resultado de este experimento ¿contradice la hipótesis habitual?

   b) Con objeto de determinar si el comportamiento agresivo citado en el apartado anterior se manifiesta del mismo modo en machos y hembras se realiza un experimento con 8 machos

y 10 hembras, observándose comportamiento agresivo en 5 machos y en 3 hembras. Calcular e interpretar un intervalo de confianza para el cociente de la proporción de machos agresivos frente a hembras agresivas.

Solución:

a) La proporción de caballas agresivas en la muestra ha resultado ser $p = \frac{4}{12} = 0{,}33$, lo cual contradice la hipótesis habitual de que $\pi \geq 0{,}60$. Ahora bien, como la muestra es pequeña, *cabe preguntarse si el resultado observado en la muestra es fruto del azar (error experimental) y no de que la hipótesis habitual sea falsa*. Para responder a esta pregunta debemos calcular cuál es *el número mínimo más probable* de caballas agresivas que cabría esperar por azar en una muestra de tamaño 12 cuando $\pi \geq 0{,}6$. Es obvio que cuanto más bajo sea el valor de π menos caballas agresivas esperaremos encontrar. Si $\pi \geq 0{,}6$, el valor más bajo posible de π es 0.6. Llamando entonces X="*Número de caballas agresivas en una muestra de tamaño 12 cuando* $\pi = 0{,}6$*"*, se tiene que $X \approx B\,(12, 0{,}6)$ y por tanto:

$$\begin{aligned} P\,(X \geq 4) &= 1 - \sum_{k=0}^{3} P\,(X = k) = 1 - \sum_{k=0}^{3} \binom{12}{k} 0{,}6^k\,(1 - 0{,}6)^{12-k} = \\ &= 1 - 0{,}4^{12} - 12 \cdot 0{,}6 \cdot 0{,}4^{11} - 66 \cdot 0{,}6^2 \cdot 0{,}4^{10} = 0{,}98473 \end{aligned}$$

Asimismo:

$$\begin{aligned} P\,(X \geq 5) &= 1 - \sum_{k=0}^{4} P\,(X = k) = 1 - \sum_{k=0}^{4} \binom{12}{k} 0{,}6^k\,(1 - 0{,}6)^{12-k} = \\ &= = 1 - 0{,}4^{12} - 12 \cdot 0{,}6 \cdot 0{,}4^{11} - 66 \cdot 0{,}6^2 \cdot 0{,}4^{10} - 220 \cdot 0{,}6^3 \cdot 0{,}4^9 = 0{,}94269 \end{aligned}$$

Si decidimos que *el número mínimo más probable que cabe esperar por azar cuando* $\pi = 0{,}60$ debe tener una probabilidad de al menos el 95 %, el número mínimo más probable es entonces el 4 (ya que para el 5 la probabilidad es inferior al 95 %). Por tanto, de acuerdo con este criterio, si bien el haber observado 4 caballas agresivas contradice la hipótesis habitual, concluimos que este valor está dentro de lo que puede ocurrir por azar siendo cierta dicha hipótesis, por lo que no hay motivo para rechazarla.

Dado el pequeño tamaño de la muestra, el intervalo de confianza adecuado para la proporción de caballas agresivas es el de Clopper-Pearson, dado por la expresión:

$$\pi \in \left[\frac{N_E}{(n - N_E + 1)F_1 + N_E}, \frac{(N_E + 1)F_2}{(n - N_E) + (N_E + 1)F_2} \right]$$

siendo N_E el número observado de "*éxitos*" (en este caso $N_{E=4}$ caballas agresivas), n es el tamaño de la muestra ($n = 12$), y:

$$\begin{aligned} F_1 &= F_{2(n-N_E+1),2N_E,\alpha/2} = F_{18,8,0{,}025} = 4{,}03 \\ F_2 &= F_{2(N_E+1),2(n-N_E),\alpha/2} = F_{10,16,0{,}025} = 2{,}99 \end{aligned}$$

Sustituyendo todos estos valores en la expresión del intervalo, obtenemos:

$$\pi \in [0{,}09933, 0{,}65142]$$

Este intervalo puede obtenerse fácilmente en R mediante (la diferencia se debe al redondeo en los valores de F_1 y F_2):

```
> binom.test(4, 12)

        Exact binomial test

data:  4 and 12
number of successes = 4, number of trials = 12, p-value =
0.3877
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.099246 0.651124
sample estimates:
probability of success
               0.33333
```

Como puede apreciarse, el intervalo de confianza indica que cabe la posibilidad de que en la población la proporción de machos sea superior al 60 % incluso aunque en la muestra haya sido sólo del 33.33 %

b) La proporciones estimadas de sujetos agresivos son, respectivamente, $p_M = \frac{5}{8} = 0{,}625$ para los machos y $p_H = \frac{3}{10} = 0{,}3$ para las hembras. El intervalo de confianza para el cociente entre la proporción de machos agresivos y la proporción de hembras agresivas en la población viene dado por:

$$\ln\left(\frac{\pi_1}{\pi_2}\right) \in \left[\ln\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \pm z_{\alpha/2}\sqrt{\frac{(1-\hat{\pi}_1)}{n_1\hat{\pi}_1} + \frac{(1-\hat{\pi}_2)}{n_2\hat{\pi}_2}}\right]$$

(si bien ha de señalarse que este intervalo es asintótico, esto es, válido sólo para muestras grandes, por lo que aquí lo aplicamos simplemente por no disponer de otra formulación para intervalos de esta clase, y debe por tanto considerarse como una burda aproximación). Sustitu-

yendo los valores de las proporciones de machos y hembras agresivos obtenemos:

$$\ln\left(\frac{\pi_M}{\pi_H}\right) \in \left[\ln\left(\frac{5/8}{3/10}\right) \pm 1{,}96\sqrt{\frac{(1-5/8)}{5}+\frac{(1-3/10)}{3}}\right] = [-0{,}35438, 1{,}8223]$$

Ahora aplicamos la función exponencial para eliminar el logaritmo, y tenemos:

$$\frac{\pi_M}{\pi_H} \in \left[e^{-0{,}35438}, e^{1{,}8223}\right] = [0{,}70161, 6{,}1862]$$

Por tanto, aunque en la muestra la proporción de machos agresivos casi duplica a la proporción de hembras agresivas, en la población podemos estar seguros al 95 % como mucho de que la proporción de machos puede ser desde 0.70161 veces la proporción de hembras (es decir, bastante menor), hasta 6.1862 veces la proporción de hembras (por tanto, bastante mayor). Así pues, no tenemos evidencia suficiente de que la proporción de machos agresivos supere a la proporción de hembras agresivas. En R calcularíamos este intervalo (de manera más simple y precisa) mediante:

```
> library(PropCIs)
> riskscoreci(5, 8, 3, 10, conf = 0.95)
data:

95 percent confidence interval:
 0.74078 6.32578
```

La diferencia entre este intervalo y el anterior se debe a que éste es más preciso que aquél, ya que la fórmula empleada, como ya hemos señalado, solo es válida para tamaños muestrales grandes. En cualquier caso, la conclusión no cambia.

6. Se han tomado muestras de sangre de individuos de la misma especie capturados en dos zonas geográficamente aisladas. A continuación se muestran las concentraciones (en ppm) de cierto enzima medidas en estos individuos:

Muestra zona 1 (24 observaciones)
76, 80, 78, 87, 77, 80, 87, 82, 76, 78, 84, 65, 84, 86, 77, 69, 85, 78, 74, 78, 72, 84, 87, 79
Muestra zona 2 (20 observaciones)
73, 86, 98, 67, 79, 78, 73, 93, 77, 84, 99, 93, 63, 70, 91, 71, 88, 93, 71, 83

Calcular media y varianza de cada muestra. Suponiendo que esta variable es normal, calcular intervalos de confianza al 95 % para el cociente de varianzas y para la diferencia de medias.

¿Existe evidencia de que el promedio de concentración de este enzima difiera de manera importante entre ambas zonas? ¿Muestran los datos evidencia de que la concentración de este enzima es más heterogénea entre los individuos de alguna de las dos zonas?

Solución:

Para calcular medias y varianzas en ambas muestras utilizamos R :

```
> z1 = c(76, 80, 78, 87, 77, 80, 87, 82, 76, 78, 84, 65,
     84, 86, 77, 69, 85, 78, 74, 78, 72, 84, 87, 79)
> z2 = c(73, 86, 98, 67, 79, 78, 73, 93, 77, 84, 99, 93,
     63, 70, 91, 71, 88, 93, 71, 83)
> mean(z1)

[1] 79.292

> mean(z2)

[1] 81.5

> sd(z1)

[1] 5.752

> sd(z2)

[1] 10.822
```

El intervalo de confianza para la diferencia de medias es:

$$\mu_1 - \mu_2 \in \left[\left(\bar{X}_1 - \bar{X}_2 \right) \pm t_{n,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

donde:

$$n = REDONDEO \left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 \frac{1}{n_1 - 1} + \left(\frac{s_2^2}{n_2} \right)^2 \frac{1}{n_2 - 1}} \right]$$

Sustituyendo los valores que hemos calculado, obtenemos que $n = 28$. Buscamos en la tabla $t_{28,0,025} = 2{,}0484$ y sustituimos en la expresión del intervalo de confianza, obteniendo finalmente:

$$\mu_1 - \mu_2 \in [-7{,}7177, 3{,}301]$$

Como este intervalo contiene una parte negativa y otra positiva, significa que carecemos de información suficiente para asegurar que la concentración media de enzimas sea mayor en un grupo que en otro. Concluimos, pues, que no se detecta diferencia significativa entre las medias de ambos grupos (Nótese, como siempre, que no aseguramos que los dos grupos tengan la misma media, sino que no podemos asegurar que sean distintas; con la información disponible las medias resultan tan parecidas que no es posible diferenciarlas).

```
> cv = var(z1)/var(z2)
> F1 = qf(0.975, 23, 19)
> F2 = qf(0.025, 23, 19)
> intvar = cv/c(F1, F2)
```

Para comparar las varianzas utilizamos el intervalo:

$$\frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1,1-\alpha/2}} \right]$$

siendo $F_{n_1-1,n_2-1,\alpha/2} = F_{23,19,0,025} = 2{,}48$ y $F_{n_1-1,n_2-1,1-\alpha/2} = F_{23,19,0,975} = \frac{1}{F_{19,23,0,025}} = \frac{1}{2{,}38} = 0{,}42017$ (los valores de las F se han hallado interpolando de manera aproximada a partir de los valores de la tabla). Se obtienen valores más precisos utilizando R :

$$F_{23,19,0,025} = \texttt{qf(0.975,23,19))} = 2{,}4648$$

$$F_{23,19,0,975} = \texttt{qf(0.025,23,19))} = 0{,}42115$$

Sustituyendo estos valores en la expresión del intervalo de confianza para el cociente de varianzas se obtiene:

$$\frac{\sigma_1^2}{\sigma_2^2} \in [0{,}11462, 0{,}67085]$$

Como este intervalo está completamente por debajo de 1, podemos estar seguros en un 95 % de que $\frac{\sigma_1^2}{\sigma_2^2} < 1$ y por tanto que $\sigma_1^2 < \sigma_2^2$. En R este intervalo puede hallarse fácilmente mediante:

```
> var.test(z1, z2)

        F test to compare two variances

data:  z1 and z2
F = 0.2825, num df = 23, denom df = 19, p-value = 0.00467
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.11462 0.67085
sample estimates:
```

```
ratio of variances
           0.28252
```

7. Con objeto de evaluar la distancia recorrida diariamente por los delfines mulares que habitan en la región macaronésica, se equipan 48 delfines (elegidos al azar en distintas manadas) con aparatos de radio que transmiten posición y distancia recorrida. A continuación se muestra la distancia total (en km.) recorrida durante un mes (30 días) por cada uno de estos delfines:

305, 389, 200, 315, 253, 304, 284, 324, 230, 215, 214, 265, 268, 301, 313, 357, 325, 404, 390 315, 411, 376, 240, 280, 313, 280, 281, 285, 339, 199, 272, 314, 275, 298, 382, 209, 257, 331 374, 203, 326, 313, 290, 319, 345, 289, 454, 247

Construye un histograma para estos datos. ¿Se aprecia normalidad?. Calcula un intervalo de confianza al 95 % para la distancia media recorrida en un mes por un delfin. Seguidamente se muestra la distancia recorrida durante cada uno de los 30 días de muestreo para uno de estos delfines:

27, 2, 1, 20, 5, 10, 6, 3, 0, 4, 2, 3, 6, 10, 1, 7, 19, 12, 4, 18, 20, 9, 5, 4, 17, 3, 3, 9, 11, 8

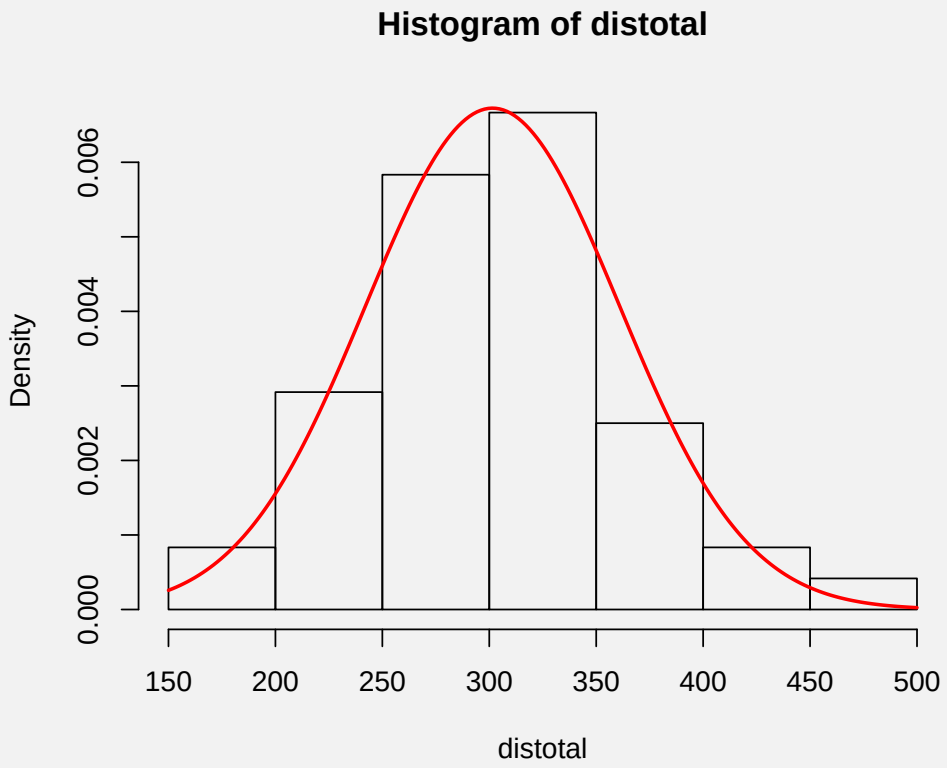
Construye un histograma para estos datos. Asumiendo que siguen una distribución exponencial, estima su parámetro y construye un intervalo de confianza al 95 % para el mismo.

Solución:

Para dibujar el histograma introducimos los datos en R ; además calculamos la media y la desviación típica y superponemos una curva normal con esos parámetros:

```
> distotal = c(305, 389, 200, 315, 253, 304, 284, 324,
     230, 215, 214, 265, 268, 301, 313, 357, 325, 404,
     390, 315, 411, 376, 240, 280, 313, 280, 281, 285,
     339, 199, 272, 314, 275, 298, 382, 209, 257, 331,
     374, 203, 326, 313, 290, 319, 345, 289, 454, 247)
> m = mean(distotal)
> m

[1] 301.52

> s = sd(distotal)
> s
```

```
[1] 59.31

> base = seq(150, 500, length = 200)
> campana = dnorm(base, m, s)
> hist(distotal, freq = F)
> lines(base, campana, col = "red", lwd = 2)
```



Como podemos observar, la curva normal se ajusta razonablemente bien al perfil del histograma, lo que sugiere que, efectivamente, estos datos siguen una distribución normal. El intervalo de confianza al 95 % para la distancia media recorrida en un mes es:

$$\begin{aligned} \mu \quad &\in \quad \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right] = \left[301{,}52 \pm t_{47,0{,}025} \frac{59{,}31}{\sqrt{48}}\right] = \\ &= \quad [284{,}3, 318{,}74] \end{aligned}$$

(de la tabla de la t de Student podemos aproximar el valor de $t_{47,0,025} \cong 2{,}012$; o bien, utilizando R $t_{47,0,025} =$ `qt(0.975,47)` $= 2{,}0117$). El intervalo de confianza para la media puede obtenerse directamente en R mediante:

```
> t.test(distotal)
```

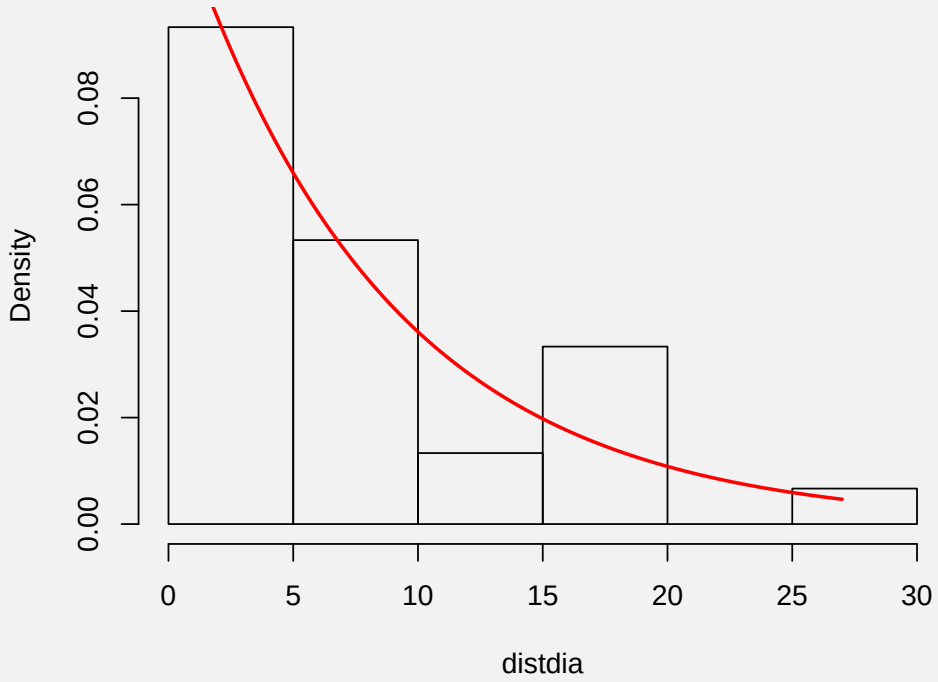
```
        One Sample t-test

data:  distotal
t = 35.221, df = 47, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 284.30 318.74
sample estimates:
mean of x
   301.52
```

Mostramos a continuación el histograma para la distancia diaria recorrida por un delfín, superponiéndole el ajuste a la distribución exponencial:

```
> distdia = c(27, 2, 1, 20, 5, 10, 6, 3, 0, 4, 2, 3, 6,
     10, 1, 7, 19, 12, 4, 18, 20, 9, 5, 4, 17, 3, 3, 9,
     11, 8)
> m = mean(distdia)
> m
```

```
[1] 8.3
```

```
> base = seq(0, max(distdia), length = 200)
> ajuste = dexp(base, 1/m)
> hist(distdia, freq = F)
> lines(base, ajuste, col = "red", lwd = 2)
```



Nuevamente vemos que el ajuste a la distribución exponencial resulta razonable. El intervalo de confianza para el parámetro λ de la distribución exponencial viene dado por:

$$\lambda \in \left[\frac{\chi^2_{2n,1-\alpha/2}}{2n\bar{X}}, \frac{\chi^2_{2n,\alpha/2}}{2n\bar{X}} \right]$$

siendo $\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{8{,}3} = 0{,}12048$ un estimador de dicho parámetro. En este caso $n = 30$, $\chi^2_{60,0,075} = 40{,}482$ y $\chi^2_{60,0,025} = 83{,}298$, y sustituyendo en la expresión anterior resulta:

$$\lambda \in [0{,}081289, 0{,}16726]$$

En R :

```
> n = 30
> qchisq(c(0.025, 0.975), 2 * n)/(2 * n * mean(distdia))

[1] 0.081289 0.167264
```

Dado que en la distribución exponencial la esperanza μ es la inversa del parámetro λ (esto es, $\mu = \frac{1}{\lambda}$), podemos calcular un intervalo para la distancia media diaria recorrida por un delfín mediante:

$$\mu \in \left[\frac{1}{0{,}16726}, \frac{1}{0{,}081289} \right] = [5{,}9786, 12{,}302]$$

Si hubiésemos utilizado la aproximación de la distribución normal, el intervalo habría sido (calculamos primero $s = \texttt{sd(distdia)} = 6{,}9289$):

$$\begin{aligned} \mu \quad &\in \quad \left[\bar{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right] = \left[8{,}3 \pm t_{29,0{,}025} \frac{6{,}9289}{\sqrt{30}}\right] = \\ &= \quad = [5{,}7127, 10{,}887] \end{aligned}$$

Como vemos, este intervalo se parece al anterior; como la muestra es de tamaño 30, ya empieza a notarse el teorema central del límite, y aunque la variable de origen no sea normal, el intervalo que presupone normalidad se parece al intervalo que realmente se ajusta a este caso. No obstante, como vemos, el intervalo basado en la normalidad tiende a subestimar el extremo superior. Como consecuencia práctica de estas observaciones resulta de interés darse cuenta de que si contamos con la información de cuál es la distribución de probabilidad de la variable que se observa, si la muestra no es demasiado grande es mejor utilizar dicha información antes que aproximar el intervalo confiando en el teorema central del límite.