ESTADÍSTICA **GRADO EN CIENCIAS DEL MAR, CURSO 2011-12**

ANÁLISIS DE DATOS - TAREA Nº 2 **INFERENCIA ESTADÍSTICA**

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

ALUMNO	
DNI:	

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

A partir de los datos contenidos en el archivo tortugas_n.csv, y cuya estadística descriptiva ya se realizó en la tarea nº 1 de este curso, responder a las cuestiones que se enuncian a continuación. Cuando para la respuesta se requiera realizar un contraste de hipótesis utilizar el nivel de significación 0.05. Si se requiere un intervalo de confianza utilizar una confianza del 95 %.

Solución:

Antes de comenzar el tratamiento de los datos, asignaremos 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. Asimismo aplicamos el comando `attach()` para poder acceder directamente a las variables contenidas en el conjunto de datos:

```
> 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)
```

1. ¿Difieren significativamente las proporciones de machos encontrados en las playas de Ervatao y Ponta Cosme? Calcula un intervalo de confianza para la proporción de machos en cada playa.

La diferencia es significativa:		p-valor:	

Intervalos de confianza	Extremo Inferior	Extremo Superior
Playa de Ervatao		
Playa de Ponta Cosme		

Solución:

El comando `table()` nos permite contar el número de ejemplares de cada sexo en cada playa:

```
> table(playa, sexo)

                sexo
playa            Hembra Macho
  Ervatao           501    57
  Ponta Cosme       198    21
  Calheta           342    31
  Porto Ferreira    259    22
```

y `prop.table()` nos proporciona la estimación de la proporción de individuos de cada sexo a partir de la tabla anterior; añadimos la opción `1` al final para que calcule la proporción dentro de cada fila (playa) de la tabla:

```
> prop.table(table(playa, sexo), 1)

                sexo
playa             Hembra   Macho
  Ervatao        0.89785 0.10215
  Ponta Cosme    0.90411 0.09589
  Calheta        0.91689 0.08311
  Porto Ferreira 0.92171 0.07829
```

Para determinar si la diferencia entre Ervatao y Ponta Cosme es significativa aplicaremos la función `prop.test()`, que lleva a cabo el contraste:

$$\begin{cases} H_0: & \pi_1 = \pi_2 \\ H_1: & \pi_0 \neq \pi_2 \end{cases}$$

siendo π_1 y π_2, respectivamente, las proporciones de machos en las dos playas. Como solamente nos interesa comparar las dos primeras playas de la tabla anterior, indicamos a R que deseamos hacer el contraste sólo con los datos de las dos primeras filas de la tabla:

```
> prop.test(table(playa, sexo)[1:2, ])

        2-sample test for equality of proportions with continuity
        correction

data:  table(playa, sexo)[1:2, ]
X-squared = 0.0165, df = 1, p-value = 0.8977
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.05583  0.04331
sample estimates:
prop 1 prop 2
0.8978 0.9041
```

Como vemos, la diferencia entre proporciones no es significativa (p-valor 0.8977, y además vemos que el intervalo para la diferencia contiene al cero). Señalemos que en realidad R ha contrastado si *la proporción de hembras* es el misma en ambas playas, ya que toma como categoría de referencia para el contraste la primera que aparece en la tabla, en este caso las hembras. Obviamente si las proporciones de hembras son iguales, también lo han de ser las proporciones de machos. En cualquier caso, si deseamos que en ese comando aparezcan explícitamente las proporciones de machos, habremos de alterar el orden de las columnas de la tabla simplemente mediante:

```
> prop.test(table(playa, sexo)[1:2, 2:1])

        2-sample test for equality of proportions with continuity
        correction

data:  table(playa, sexo)[1:2, 2:1]
X-squared = 0.0165, df = 1, p-value = 0.8977
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
 -0.04331  0.05583
sample estimates:
 prop 1  prop 2
0.10215 0.09589
```

Por último, para calcular un intervalo de confianza para la proporción de machos en cada playa utilizamos `binom.test()`. Podemos introducir directamente el número de machos y el número de ejemplares en cada playa que hemos obtenido más arriba, o aplicar esta función directamente a la tabla de frecuencias:

```
> binom.test(table(sexo[playa == "Ervatao"]))

        Exact binomial test

data:  table(sexo[playa == "Ervatao"])
number of successes = 501, number of trials = 558, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.8697 0.9217
sample estimates:
probability of success
                0.8978
```

Vemos que, por defecto, se ha calculado el intervalo de confianza para la primera categoría de la variable `sexo`, en este caso "*Hembra*" ya que se usa el orden alfabético. Para que calcule el intervalo para los machos, invertimos el orden de los términos en la tabla para que la primera categoría sea la de los machos:

```
> binom.test(table(sexo[playa == "Ervatao"])[2:1])

        Exact binomial test

data:  table(sexo[playa == "Ervatao"])[2:1]
number of successes = 57, number of trials = 558, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.07829 0.13032
sample estimates:
```

```
probability of success
               0.1022

> binom.test(table(sexo[playa == "Ponta Cosme"])[2:1])

        Exact binomial test

data:  table(sexo[playa == "Ponta Cosme"])[2:1]
number of successes = 21, number of trials = 219, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.06034 0.14284
sample estimates:
probability of success
              0.09589
```

Respuesta correcta:

La diferencia es significativa:		p-valor:	
Intervalos de confianza	Extremo Inferior		Extremo Superior
Playa de Ervatao			
Playa de Ponta Cosme			

2. ¿Podría asegurarse que las hembras son mayores que los machos en alguna de las variables LCC, ACC o peso? Calcula en cada caso un intervalo de confianza al 95 % para la diferencia media entre sexos.

Variable	¿Son las hembras, en media, mayores que los machos?	p-valor:	Intervalo de Confianza	
			Ext. Inf.	Ext. Sup.
LCC				
ACC				
Peso				

Solución:

Para este estudio se dispone de una muestra grande de ejemplares de cada sexo, (si bien es cierto que hay muchas más hembras que machos):

```
> table(sexo)

sexo
Hembra  Macho
  1300    131
```

Por esta razón no nos preocupamos de contrastar la normalidad de las variables y procedemos directamente con `t.test()` para contrastar las diferencias de medias y construir los intervalos de confianza. Los tres contrastes solicitados son de la forma:

$$\begin{cases} H_0: & \mu_H \leq \mu_M \\ H_1: & \mu_H > \mu_M \end{cases}$$

(hemos puesto la media correspondiente a las hembras en primer lugar ya que, por defecto, R utiliza el orden alfabético). Por tanto la sintaxis a emplear es:

```
> t.test(LCC ~ sexo, alternative = "greater")

        Welch Two Sample t-test


data:  LCC by sexo
t = 4.061, df = 194.5, p-value = 3.541e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6712    Inf
sample estimates:
mean in group Hembra  mean in group Macho
               82.00                80.86

> t.test(ACC ~ sexo, alternative = "greater")

        Welch Two Sample t-test


data:  ACC by sexo
t = 0.2576, df = 148.9, p-value = 0.3985
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.5708     Inf
sample estimates:
mean in group Hembra  mean in group Macho
               77.05                76.95

> t.test(peso ~ sexo, alternative = "greater")
```

```
        Welch Two Sample t-test

data:  peso by sexo
t = 1.331, df = 160.5, p-value = 0.09252
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1406     Inf
sample estimates:
mean in group Hembra  mean in group Macho
               60.80                60.22
```

Para LCC el p-valor es 3.541e-05, menor que 0.05, por lo que concluimos que existe evidencia suficiente de que las hembras tienen, en media, una mayor LCC que los machos. En el caso de ACC, el p-valor es 0.3985, mayor que 0.05, por lo que concluimos que no existe evidencia de que las hembras tengan, en media, una mayor ACC que los machos. Por último, con respecto al peso, el p-valor es 0.09252, mayor que 0.05, por lo que concluimos que no existe evidencia de que las hembras tengan, en media, un peso mayor que los machos.

Por último, para calcular los intervalos de confianza para las diferencias de medias volvemos a utilizar `t.test()` sin indicar ahora ninguna hipótesis alternativa:

```
> t.test(LCC ~ sexo)$conf.int

[1] 0.5821 1.6815
attr(,"conf.level")
[1] 0.95

> t.test(ACC ~ sexo)$conf.int

[1] -0.7018  0.9123
attr(,"conf.level")
[1] 0.95

> t.test(peso ~ sexo)$conf.int

[1] -0.2799  1.4376
attr(,"conf.level")
[1] 0.95
```

Respuesta correcta:

Variable	¿Son las hembras, en media, mayores que los machos?	p-valor:	Intervalo de Confianza	
			Ext. Inf.	Ext. Sup.
LCC				
ACC				
Peso				

3. Estima el valor medio de LCC en hembras y calcula el error estándar de dicha estimación[1]. Se ha documentado que la Longitud Curva del Caparazón en las hembras de esta especie es, en media, inferior a 80 cm. Con los datos de esta muestra ¿hay evidencia suficiente para afirmar que en realidad se supera este valor?

Valor medio observado de LCC en hembras.	Error estándar de la media	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.

¿Hay evidencia suficiente de que en las hembras ocurre que $\mu_{LCC} > 80$?	p-valor:

Solución:
Nuevamente, como en el caso anterior, la muestra es lo suficientemente grande como para no tener que preocuparnos de la normalidad de la variable. Podemos calcular la LCC media, así como su error estándar $(s/\sqrt{n})$ mediante:

```
> mean(LCC[sexo == "Hembra"])

[1] 82

> sd(LCC[sexo == "Hembra"])/sqrt(length(!is.na(LCC[sexo ==
    "Hembra"])))

[1] 0.1198
```

Hemos empleado el comando `length(!is.na(LCC[sexo=="Hembra"]))` para contar el número de valores no perdidos (`!is.na()`) que contiene la variable LCC medida sobre las hembras.

Podemos calcular el intervalo de confianza mediante:

[1] Recuerda que se define el error estándar de la media como la desviación típica de la variable dividida por la raíz cuadrada del tamaño de la muestra.

```
> t.test(LCC[sexo == "Hembra"])$conf.int

[1] 81.76 82.23
attr(,"conf.level")
[1] 0.95
```

y contrastar si el valor medio de esta variable es mayor que 80 mediante `t.test()`:

```
> t.test(LCC[sexo == "Hembra"], mu = 80, alternative = "greater")

        One Sample t-test

data:  LCC[sexo == "Hembra"]
t = 16.67, df = 1299, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 80
95 percent confidence interval:
 81.8  Inf
sample estimates:
mean of x
       82
```

Como el p-valor es menor que 0.05, concluimos que existe evidencia suficiente de que las hembras tienen, en media, una LCC mayor que 80 cm.

Respuesta correcta:

Valor medio observado de LCC en hembras.	Error estándar de la media	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.

¿Hay evidencia suficiente de que en las hembras ocurre que $\mu_{LCC} > 80$?	p-valor:

4. ¿Existe evidencia suficiente de que las hembras ponen, por término medio, más huevos en el primer periodo (16-31 de julio) que en el quinto (16-30 de septiembre)? Calcula un intervalo de confianza al 95 % para la diferencia media en el tamaño de la puesta entre ambos periodos.

¿Ponen las hembras por término medio más huevos en el primer periodo que en el quinto?	p-valor:

Diferencia media en el tamaño de la puesta (periodo 1 - periodo 5)	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.

Solución:

En primer lugar contamos el número de observaciones por periodo:

```
> table(periodo)

periodo
     del 16 al 31 de julio       del 1 al 15 de agosto
                       298                         313
    del 16 al 31 de agosto  del 1 al 15 de septiembre
                       267                         274
del 16 al 30 de septiembre
                       279
```

El número de observaciones en cada periodo es lo suficiente grande como para poder aproximar el tamaño medio de la puesta mediante una distribución normal y poder aplicar, por tanto, el `t.test()` para comparar las medias:

```
> t.test(Huevos ~ periodo, data = subset(tortugas, periodo ==
     "del 16 al 31 de julio" | periodo == "del 16 al 30 de septiembre"),
     alternative = "greater")

        Welch Two Sample t-test

data:  Huevos by periodo
t = 9.811, df = 501.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 8.512   Inf
sample estimates:
     mean in group del 16 al 31 de julio
                                   86.63
mean in group del 16 al 30 de septiembre
                                   76.40
```

Como el p-valor es menor que 0.05, concluimos que existe evidencia suficiente de que las hembras ponen, en media, más huevos en el primer periodo que en el último.

Para calcular la diferencia de medias y obtener un intervalo de confianza repetimos el `t.test()` anterior sin especificar hipótesis alternativa:

```
> t.test(Huevos ~ periodo, data = subset(tortugas, periodo ==
     "del 16 al 31 de julio" | periodo == "del 16 al 30 de septiembre"))

        Welch Two Sample t-test


data:  Huevos by periodo
t = 9.811, df = 501.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.181 12.279
sample estimates:
     mean in group del 16 al 31 de julio
                                   86.63
mean in group del 16 al 30 de septiembre
                                   76.40
```

La diferencia en el tamaño de las puestas entre ambos periodos es, pues, de $86{,}63 - 76{,}4 = 10{,}23$ huevos en promedio, con un intervalo de confianza al 95 % dado por $[8{,}181, 12{,}28]$.

Respuesta correcta:

¿Ponen las hembras por término medio más huevos en el primer periodo que en el quinto?	p-valor:

Diferencia media en el tamaño de la puesta (periodo 1 - periodo 5)	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.

5. Estima la profundidad media a que se excavan los nidos. Estima también la distancia media desde los nidos a la línea de marea. Acompaña ambas estimaciones de sendos intervalos de confianza al 95 %.

Variable	Media	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
Profundidad Nido			
Distancia a Linea de Marea			

Solución:

Utilizamos `t.test()` para ambas estimaciones, ya que el gran número de observaciones permite utilizar la aproximación normal:

```
> t.test(profNido)

        One Sample t-test

data:  profNido
t = 300, df = 1299, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.65 57.40
sample estimates:
mean of x
    57.02

> t.test(distancia)

        One Sample t-test

data:  distancia
t = 111.6, df = 1299, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 23.79 24.64
sample estimates:
mean of x
    24.21
```

Respuesta correcta:

Variable	Media	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
Profundidad Nido			
Distancia a Linea de Marea			

6. Estima la proporción de nidos en los que se encuentran galería excavadas por cangrejos, y acompaña dicha estimación de un intervalo de confianza al 95 %. ¿Se puede asegurar que más de la mitad de los nidos presentan estas galerías?

Proporción estimada	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.

¿Existe evidencia suficiente de que más de la mitad de los nidos presentan galerías de cangrejo?	p-valor:

Solución:

Contamos en primer lugar cuántos nidos tienen galerías excavadas por cangrejos y cuantos no:

```
> table(cangrejos)

cangrejos
 NO  SÍ
365 573
```

Para estimar la proporción de nidos con galerías de cangrejos utilizamos `prop.table()`:

```
> prop.table(table(cangrejos))

cangrejos
    NO     SÍ
0.3891 0.6109
```

Ahora, para contrastar si hay suficiente evidencia de que los excavados por cangrejos son más de la mitad utilizamos `binom.test()` (cambiando el orden de los términos en la tabla, para que los nidos excavados sean contabilizados en primer lugar):

```
> binom.test(table(cangrejos)[2:1], p = 0.5, alternative = "greater")

        Exact binomial test

data:  table(cangrejos)[2:1]
number of successes = 573, number of trials = 938, p-value =
5.76e-12
alternative hypothesis: true probability of success is greater than 0.5
```

```
95 percent confidence interval:
 0.5839 1.0000
sample estimates:
probability of success
                0.6109
```

Como el p-valor es menor que 0.05, concluimos que existe evidencia suficiente de que los nidos excavados por cangrejos son efectivamente más de la mitad.

Utilizamos de nuevo `binom.test()`, pero ahora sin especificar alternativa, para obtener un intervalo de confianza para esta proporción:

```
> binom.test(table(cangrejos)[2:1])$conf.int
```

```
[1] 0.5788 0.6422
attr(,"conf.level")
[1] 0.95
```

Respuesta correcta:

Proporción estimada	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.

¿Existe evidencia suficiente de que más de la mitad de los nidos presentan galerías de cangrejo?	p-valor:

7. Estima la recta de regresión para predecir la profundidad del nido a partir de la distancia a la linea de marea. ¿Se puede asegurar que a mayor distancia de dicha linea es mayor la profundidad del nido? Estima mediante un intervalo de confianza en cuánto aumenta la profundidad del nido por cada metro que la posición de éste se aleja de la linea de marea. Estima también, mediante un intervalo de confianza la profundidad media que cabe esperar en los nidos situados a 25 metros de dicha linea.

Recta de regresión	Pendiente:		Ordenada:	

¿Se puede asegurar que a mayor distancia de la linea de marea es mayor la profundidad del nido?	p-valor:

Variable	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.
Incremento en prof. por cada metro adicional desde la linea de marea:		
Profundidad media en nidos a 25 m. de la línea de marea:		

Solución:

La estimación de la recta de regresión se realiza mediante el comando `lm()`:

```
> regre = lm(profNido ~ distancia)
> regre

Call:
lm(formula = profNido ~ distancia)

Coefficients:
(Intercept)    distancia
     43.040        0.578
```

La pendiente de esta regresión (0.5776) nos indica en cuanto aumenta la profundidad del nido (en cm.) por cada metro adicional que éste se encuentre alejado de la línea de marea. Para determinar si puede asegurarse que a mayor distancia se excavan efectivamente nidos más profundos hemos de decidir si el valor de esta pendiente es significativo. Utilizando `summary()` obtenemos una estimación más completa de los elementos de la regresión:

```
> summary(regre)

Call:
lm(formula = profNido ~ distancia)

Residuals:
    Min      1Q  Median      3Q     Max
-15.408  -3.482  -0.018   3.556  20.021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.0402     0.4654    92.5   <2e-16 ***
distancia     0.5776     0.0183    31.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.16 on 1298 degrees of freedom
  (131 observations deleted due to missingness)
Multiple R-squared: 0.434,        Adjusted R-squared: 0.434
F-statistic:  997 on 1 and 1298 DF,  p-value: <2e-16
```

En particular, el último p-valor mostrado corresponde al test F asociado al contraste de la regresión:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

Como este p-valor es menor que 0.05, concluimos que existe evidencia suficiente de que la pendiente es distinta de cero. La siguiente sintaxis calcula intervalos de confianza para la ordenada y la pendiente de la recta:

```
> confint(regre)
```

```
              2.5 %  97.5 %
(Intercept) 42.1273 43.9531
distancia    0.5417  0.6134
```

La primera fila de esta tabla es el intervalo para la ordenada y la segunda fila es el intervalo para la pendiente. Este último intervalo es íntegramente positivo, lo que indica que existe fuerte evidencia de que la pendiente es positiva y por tanto, podemos asegurar que a medida que aumenta la distancia a la línea de marea aumenta la profundidad de los nidos.

Por último, la predicción de la profundidad media en nidos situados a 25 metros de la línea de marea podemos llevarla a cabo del siguiente modo:

```
> datapred = data.frame(distancia = 25)
> predict(regre, datapred, interval = "confidence")
```

```
    fit  lwr   upr
1 57.48 57.2 57.76
```

Respuesta correcta:

Recta de regresión	Pendiente:		Ordenada:	

¿Se puede asegurar que a mayor distancia de la linea de marea es mayor la profundidad del nido?	p-valor:

Variable	Intervalo de Confianza	
	Ext. Inf.	Ext. Sup.
Incremento en prof. por cada metro adicional desde la linea de marea:		
Profundidad media en nidos a 25 m. de la línea de marea:		

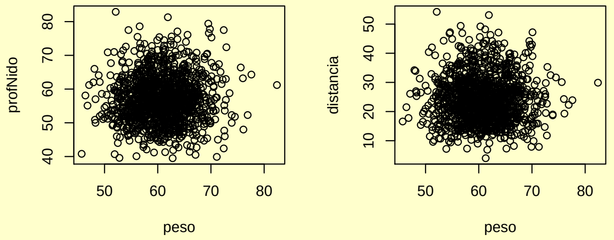
8. ¿Existe asociación lineal entre el peso o LCC de la tortuga y la profundidad o distancia a que se excava el nido?

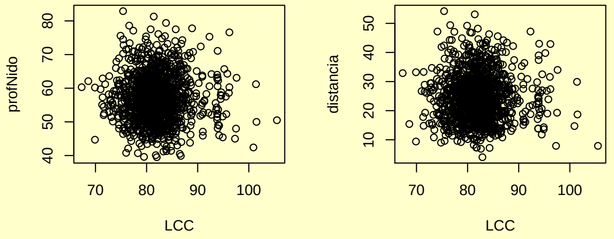
x	y	Pendiente	Ordenada	¿Existe asociación?	p-valor
Peso	Profundidad				
Peso	Distancia				
LCC	Profundidad				
LCC	Distancia				

Solución:

En la siguiente figura se representan gráficamente las nubes de puntos correspondientes a las variables indicadas:

```
> par(mfrow = c(2, 2))
> plot(peso, profNido)
> plot(peso, distancia)
> plot(LCC, profNido)
> plot(LCC, distancia)
```





Los gráficos muestran que entre estas variables no hay asociación lineal o, en caso de haberla, es muy débil. Para estimar las regresiones empleamos nuevamente el comando `lm()`:

```
> lm(profNido ~ peso)

Call:
lm(formula = profNido ~ peso)


Coefficients:
(Intercept)         peso
    53.7520       0.0538

> lm(distancia ~ peso)

Call:
lm(formula = distancia ~ peso)


Coefficients:
(Intercept)         peso
   24.59234     -0.00624

> lm(profNido ~ LCC)

Call:
lm(formula = profNido ~ LCC)


Coefficients:
(Intercept)          LCC
     56.287        0.009
```

```
> lm(distancia ~ LCC)

Call:
lm(formula = distancia ~ LCC)

Coefficients:
(Intercept)          LCC
   23.44974      0.00931
```

Y para determinar si la asociación es significativa en cada caso podemos proceder como en el problema anterior, haciendo el contraste de regresión, y comparando el p-valor del test F con nuestro nivel de significación α (que fijamos en 0.05). Sólo si el p-valor es menor que α concluimos que la asociación es significativa:

```
> summary(lm(profNido ~ peso))

Call:
lm(formula = profNido ~ peso)

Residuals:
    Min      1Q  Median      3Q     Max
-17.679  -4.769  -0.469   4.256  26.343

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.7520     2.3332   23.04   <2e-16 ***
peso          0.0538     0.0382    1.41     0.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.85 on 1298 degrees of freedom
  (131 observations deleted due to missingness)
Multiple R-squared: 0.00152,       Adjusted R-squared: 0.000754
F-statistic: 1.98 on 1 and 1298 DF,  p-value: 0.16

> summary(lm(distancia ~ peso))

Call:
lm(formula = distancia ~ peso)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-20.210  -5.747  -0.696   4.736  29.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.59234    2.66481    9.23   <2e-16 ***
peso        -0.00624    0.04368   -0.14     0.89
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.82 on 1298 degrees of freedom
  (131 observations deleted due to missingness)
Multiple R-squared: 1.57e-05,        Adjusted R-squared: -0.000755
F-statistic: 0.0204 on 1 and 1298 DF,  p-value: 0.886

> summary(lm(profNido ~ LCC))

Call:
lm(formula = profNido ~ LCC)

Residuals:
    Min      1Q  Median      3Q     Max
-17.525  -4.800  -0.443   4.256  25.935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   56.287      3.617    15.6   <2e-16 ***
LCC            0.009      0.044     0.2     0.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.86 on 1298 degrees of freedom
  (131 observations deleted due to missingness)
Multiple R-squared: 3.22e-05,        Adjusted R-squared: -0.000738
F-statistic: 0.0418 on 1 and 1298 DF,  p-value: 0.838

> summary(lm(distancia ~ LCC))

Call:
lm(formula = distancia ~ LCC)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-20.222  -5.690  -0.713   4.726  30.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.44974    4.12819    5.68  1.7e-08 ***
LCC          0.00931    0.05028    0.19     0.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.82 on 1298 degrees of freedom
  (131 observations deleted due to missingness)
Multiple R-squared: 2.64e-05,       Adjusted R-squared: -0.000744
F-statistic: 0.0343 on 1 and 1298 DF,  p-value: 0.853
```

Respuesta correcta:

x	y	Pendiente	Ordenada	¿Existe asociación?	p-valor
Peso	Profundidad				
Peso	Distancia				
LCC	Profundidad				
LCC	Distancia				

9. Se sospecha que la distancia media desde la posición del nido a la línea de marea puede variar entre playas, debido fundamentalemente a la pendiente de las mismas, la proximidad de accidentes geográficos y la exposición al viento.

   a) Utilizando el análisis de la varianza determina si existen diferencias significativas en la distancia media desde los nidos a la linea de marea entre las distintas playas. Valida la realización del análisis de la varianza verificando las hipótesis de normalidad y homoscedasticidad.

Cuestión	Respuesta	p-valor
¿Existen diferencias significativas entre playas?		
¿Los residuos del ANOVA son normales? (test de Shapiro-Wilk)		
¿Se verifica la hipótesis de homoscedasticidad? (test de Levene)		
¿Es válida la aplicación del ANOVA?		

Solución:

Podemos ajustar el modelo de análisis de la varianza mediante:

```
> adeva = aov(distancia ~ playa)
```

El comando `model.tables()` nos permite ver cuál es la distancia media de los nidos a la línea de marea en cada playa, mientras que `boxplot()` nos representa gráficamente la distribución de las distancias por playa:

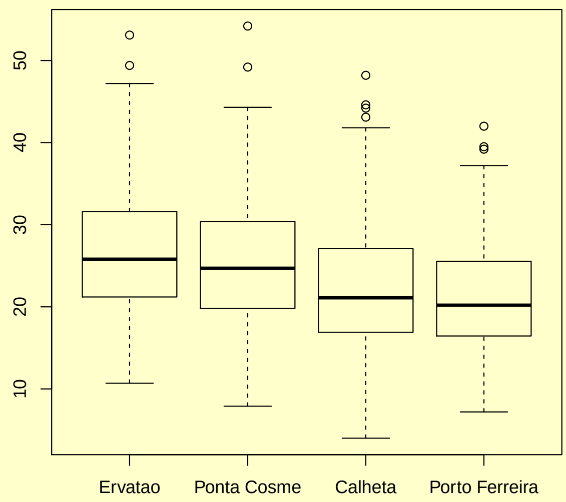
```
> model.tables(adeva, "means")

Tables of means
Grand mean

24.21

 playa
    Ervatao Ponta Cosme Calheta Porto Ferreira
      26.58       25.83   22.16          21.11
rep  501.00      198.00  342.00         259.00

> boxplot(distancia ~ playa)
```



Así pues, podemos apreciar que la distancia media de los nidos a la línea de marea varía desde los 21.11 metros en la playa de Porto Ferreira hasta 26.58 metros en la playa de Ervatao. Para determinar si las diferencias entre playas son significativas calculamos el test F y su p-valor:

```
> summary(adeva)

              Df Sum Sq Mean Sq F value Pr(>F)
playa          3   7249    2416    43.4 <2e-16 ***
Residuals   1296  72216      56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
131 observations deleted due to missingness
```

El p-valor es menor que 0.05 (valor que elegimos como nivel de significación), por lo que concluimos que existe evidencia suficiente de que la distancia media entre los nidos y la línea de marea varía de manera significativa entre playas.

Para verificar la validez de este análisis deben comprobarse las hipótesis de normalidad y homoscedasticidad. La normalidad la comprobamos mediante el test de Shapiro-Wilk:

```
> shapiro.test(residuals(adeva))

        Shapiro-Wilk normality test


data:  residuals(adeva)
W = 0.9815, p-value = 7.24e-12
```

El p-valor de este contraste es menor que 0.05, por lo que no podemos aceptar la normalidad de los residuos.

Para comprobar la homoscedasticidad utilizamos el test de Levene:

```
> library(car)
> leveneTest(distancia ~ playa)

Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    3    2.86  0.036 *
      1296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En este contraste el p-valor es menor que 0.05, por lo que no podemos aceptar la homoscedasticidad de la variable *distancia* en las distintas playas.

Por tanto, como no puede aceptarse la hipótesis de normalidad, y tampoco puede aceptarse la hipótesis de homoscedasticidad, la aplicación del análisis de la varianza no resulta adecuada con estos datos.

Respuesta correcta:

Cuestión	Respuesta	p-valor
¿Existen diferencias significativas entre playas?		
¿Los residuos del ANOVA son normales? (test de Shapiro-Wilk)		
¿Se verifica la hipótesis de homoscedasticidad? (test de Levene)		
¿Es válida la aplicación del ANOVA?		

b) Repite el apartado anterior considerando como variable respuesta la raiz cuadrada de la distancia desde el nido a la línea de marea.

Cuestión	Respuesta	p-valor
¿Existen diferencias significativas entre playas?		
¿Los residuos del ANOVA son normales? (test de Shapiro-Wilk)		
¿Se verifica la hipótesis de homoscedasticidad? (test de Levene)		
¿Es válida la aplicación del ANOVA?		

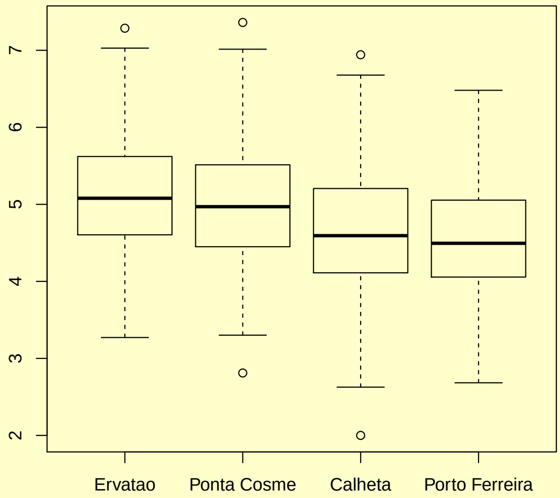
Solución:

Definimos la nueva variable `raizDist` como la raiz cuadrada de la distancia desde la posición del nido a la línea de la marea, y ajustamos el modelo de análisis de la varianza con esta nueva variable mediante:

```
> raizDist = sqrt(distancia)
> adeva = aov(raizDist ~ playa)
```

Utilizamos `boxplot()` para representar gráficamente la distribución de la nueva variable en las distintas playas:

```
> boxplot(raizDist ~ playa)
```



Observamos ahora unas distribuciones más simétricas y con menos valores extremos. Para determinar si las diferencias entre playas, medidas de acuerdo a la variable `raizDist`, son significativas calculamos el test F y su p-valor:

```
> summary(adeva)

              Df Sum Sq Mean Sq F value Pr(>F)
playa          3     77   25.76    45.1 <2e-16 ***
Residuals   1296    741    0.57
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
131 observations deleted due to missingness
```

El p-valor es menor que 0.05 (valor que elegimos como nivel de significación), por lo que concluimos que existe evidencia suficiente de que el valor medio de la raiz cuadrada de la distancia entre los nidos y la línea de marea varía de manera significativa entre playas.

Para verificar la validez de este análisis comprobamos las hipótesis de normalidad y homoscedasticidad. Comprobamos la normalidad mediante el test de Shapiro-Wilk:

```
> shapiro.test(residuals(adeva))

        Shapiro-Wilk normality test

data:  residuals(adeva)
W = 0.9981, p-value = 0.1328
```

El p-valor de este contraste es mayor que 0.05, por lo que aceptamos la normalidad de los residuos.

Para comprobar la homoscedasticidad utilizamos el test de Levene:

```
> library(car)
> leveneTest(raizDist ~ playa)

Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    3    1.17   0.32
      1296
```

En este contraste el p-valor es mayor que 0.05, por lo que aceptamos la homoscedasticidad de la variable *raizDist* en las distintas playas.

Por tanto, como puede aceptarse la hipótesis de normalidad, y también puede aceptarse la hipótesis de homoscedasticidad, la aplicación del análisis de la varianza resulta adecuada con estos datos.

Una nota sobre la interpretación de este análisis: en realidad nuestro interés se centra en comparar la *distancia* desde los nidos a la línea de marea alta en las distintas playas. Tal como se ha visto en el primer apartado de esta pregunta la estrategia de aplicar el análisis de la varianza no vale con la variable *distancia* original ya que no se cumplen los supuestos sobre los que se apoya dicho método; no obstante, el cambio de escala (cambiar la distancia por su raíz cuadrada) consigue que tales supuestos sí que se cumplan, por lo que el análisis de la varianza en este caso sí que resulta válido. Por tanto, si bien lo que hemos detectado en realidad es que existen diferencias significativas entre los valores medios de *las raíces cuadradas de las distancias* de los nidos a la linea de marea alta, ello en el fondo constituye una evidencia indirecta de que las distancias originales deben ser también distintas entre playas.

Respuesta correcta:

Cuestión	Respuesta	p-valor
¿Existen diferencias significativas entre playas?		
¿Los residuos del ANOVA son normales? (test de Shapiro-Wilk)		
¿Se verifica la hipótesis de homoscedasticidad? (test de Levene)		
¿Es válida la aplicación del ANOVA?		

10. El éxito de emergencia (definido como el cociente entre el número de crías emergidas y el tamaño de la puesta) se considera *Alto* si es superior a 0.2 y *Bajo* si es inferior a esta cantidad. Estima la

proporción de nidos con alto éxito de emergencia en cada playa y determina, utilizando el test de la chi-cuadrado, si dicha proporción presenta diferencias significativas entre las distintas playas.

Playa	Proporción	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
Ervatao			
Ponta Cosme			
Calheta			
Porto Ferreira			

¿Difiere esta proporción significativamente entre playas?		pvalor:	

Solución:

Calculamos en primer lugar el éxito de emergencia de cada nido:

```
> exitoEmerg = crias_Emerg/Huevos
```

A partir de esta variable construimos otra que especifique si el éxito de emergencia es alto o bajo, y exploramos la proporción de nidos de cada clase a nivel global:

```
> clasifExitoEmerg = ifelse(exitoEmerg > 0.2, "Alto", "Bajo")
> table(clasifExitoEmerg)
```

```
clasifExitoEmerg
Alto Bajo
 544  394
```

```
> prop.table(table(clasifExitoEmerg))
```

```
clasifExitoEmerg
Alto Bajo
0.58 0.42
```

Así pues, considerada globalmente, en esta isla el 58 % de los nidos de tortuga presentan un alto éxito de emergencia.

Calculamos ahora la proporción de nidos con alto y bajo éxito de emergencia en cada playa:

```
> table(playa, clasifExitoEmerg)
```

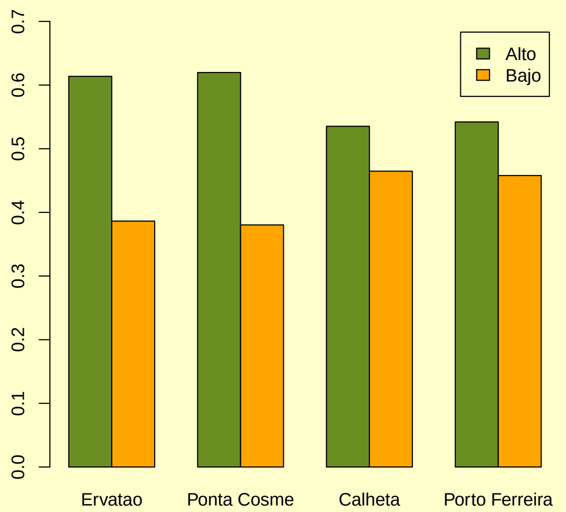
```
                clasifExitoEmerg
playa             Alto Bajo
  Ervatao          224  141
  Ponta Cosme       88   54
  Calheta          129  112
  Porto Ferreira   103   87
```

```
> prop.table(table(playa, clasifExitoEmerg), 1)

                 clasifExitoEmerg
playa               Alto   Bajo
  Ervatao         0.6137 0.3863
  Ponta Cosme     0.6197 0.3803
  Calheta         0.5353 0.4647
  Porto Ferreira 0.5421 0.4579
```

Vemos que entre las playas hay algunas diferencias; la proporción de nidos con alto éxito de emergencia oscila desde el 53.53 % de Calheta hasta el 61.97 % de Ponta Cosme. El siguiente gráfico permite visualizar las diferencias entre playas:

```
> barplot(prop.table(table(clasifExitoEmerg, playa), 2),
     beside = T, legend.text = T, col = c("olivedrab",
         "orange"), ylim = c(0, 0.7))
```



Para calcular un intervalo de confianza para cada una de las proporciones aplicamos la función `binom.test()` a cada una de las filas de la tabla en la que se cuenta el número de nidos con alto y bajo éxito de emergencia en cada playa:

```
> binom.test(table(playa, clasifExitoEmerg)[1, ])$conf.int

[1] 0.5616 0.6639
attr(,"conf.level")
[1] 0.95
```

```
> binom.test(table(playa, clasifExitoEmerg)[2, ])$conf.int

[1] 0.5345 0.6998
attr(,"conf.level")
[1] 0.95

> binom.test(table(playa, clasifExitoEmerg)[3, ])$conf.int

[1] 0.4701 0.5995
attr(,"conf.level")
[1] 0.95

> binom.test(table(playa, clasifExitoEmerg)[4, ])$conf.int

[1] 0.4685 0.6144
attr(,"conf.level")
[1] 0.95
```

Una forma más elegante (y limpia) de resolver esta tarea habría sido simplemente definir una función que extraiga el intervalo de confianza calculado por `binom.test()`, y aplicar dicha función a la misma tabla de antes:

```
> bt.ci = function(x) binom.test(x)$conf.int
> apply(table(playa, clasifExitoEmerg), 1, bt.ci)

     playa
      Ervatao Ponta Cosme Calheta Porto Ferreira
  [1,]  0.5616      0.5345  0.4701         0.4685
  [2,]  0.6639      0.6998  0.5995         0.6144
```

Por último, para determinar si las diferencias entre las proporciones de nidos con alto éxito de emergencia en las distintas playas son significativas o, por el contrario, pueden atribuirse simplemente al azar (y por tanto las proporciones poblacionales pueden considerarse idénticas), llevamos a cabo el test de la chi-cuadrado. Las hipótesis a contrastar son:

$$\begin{cases} H_0: & \pi_{Ervatao} = \pi_{Ponta\,Cosme} = \pi_{Calheta} = \pi_{Porto\,Ferreira} \\ H_1: & \text{Al menos dos proporciones son distintas entre sí} \end{cases}$$

siendo π_{Playa} la proporción de nidos con alto éxito reproductivo en la playa especificada. Este contraste puede hacerse en R de dos maneras: utilizando `chisq.test()`:

```
> chisq.test(table(playa, clasifExitoEmerg))
```

```
        Pearson's Chi-squared test

data:  table(playa, clasifExitoEmerg)
X-squared = 5.72, df = 3, p-value = 0.126
```

o utilizando `prop.test()`, que nos muestra el test anterior y añade la estimación de la proporción en cada playa:

```
> prop.test(table(playa, clasifExitoEmerg))

        4-sample test for equality of proportions without continuity
        correction

data:  table(playa, clasifExitoEmerg)
X-squared = 5.72, df = 3, p-value = 0.126
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4
0.6137 0.6197 0.5353 0.5421
```

El p-valor del test (0.126) es mayor que $\alpha = 0{,}05$, lo que indica que podemos aceptar la hipótesis nula. Por tanto las proporciones de nidos con alto éxito reproductivo no difieren de manera significativa entre playas.

Respuesta correcta:

Playa	Proporción	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
Ervatao			
Ponta Cosme			
Calheta			
Porto Ferreira			

¿Difiere esta proporción significativamente entre playas?		pvalor:	

11. De acuerdo con estudios anteriores a los años cincuenta, el 40 % de los anidamientos se producían en la playa A, el 25 % en la B, el 30 % en la C y el 5 % en otras playas. ¿Se conservan estas proporciones en la actualidad?

Respuesta:		pvalor:	

Solución:
En primer lugar calculamos el porcentaje de anidamientos producido en cada playa. Como los anidamientos corresponden sólo a las hembras, basta contar cuántas hembras se han encontrado en cada playa:

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

     Ervatao    Ponta Cosme        Calheta Porto Ferreira
         501            198            342            259

> prop.table(table(playa[sexo == "Hembra"]))

     Ervatao    Ponta Cosme        Calheta Porto Ferreira
      0.3854         0.1523         0.2631         0.1992
```

Obviamente las proporciones no coinciden exactamente con las que se observaban en los años 50 (Ervatao y Calheta son parecidas, pero Ponta Cosme y Porto Ferreira aparentan bastante distintas a como eran). La pregunta que se plantea en este caso es si estas diferencias pueden explicarse simplemente por efecto del azar o si, por el contrario, son tan grandes que necesariamente implican que ha habido un cambio real en dichas proporciones con el transcurso del tiempo. En términos de un contraste de hipótesis, podemos expresar esta cuestión de la siguiente forma:

$$\begin{cases} H_0: & \pi_{Ervatao} = 0{,}4, \;\; \pi_{Ponta\,Cosme} = 0{,}25, \;\; \pi_{Calheta} = 0{,}30, \;\; \pi_{Porto\,Ferreira} = 0{,}05 \\ H_1: & \text{Al menos una de las igualdades anteriores es falsa.} \end{cases}$$

(siendo π_{Playa} el porcentaje de anidamientos en la actualidad en cada playa). Este es un contraste de bondad de ajuste, que se resuelve empleando el test de la chi-cuadrado. El fundamento de este test es comparar las frecuencias observadas con las esperadas si fuera cierta la hipótesis nula, y calcular la discrepancia entre ambas mediante el estadístico:

$$\chi^2 = \sum_{i=1}^{4} \frac{(f_{obs_i} - f_{esp_i})^2}{f_{esp_i}}$$

Obviamente este estadístico tomará un valor pequeño si lo observado es similar a lo esperado, y un valor grande si lo que se observa está lejos de lo que se espera cuando es cierta H_0; un valor grande de este estadístico, por tanto, conducirá al rechazo de H_0. En este caso las frecuencias observadas son las que se muestran en la primera de las tablas de más arriba. Para calcular las frecuencias esperadas hemos de tener en cuenta que el número total de nidos observados fue:

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

[1] 1300
```

Las frecuencias esperadas f_{esp_i} en cada playa se calculan simplemente multiplicando este número por las proporciones teóricas (0.4, 0.25, 0.3 y 0.05), por lo que el número de nidos esperado en cada playa si fuera cierta la hipótesis nula (que la distribución no ha cambiado desde los años 50) sería:

```
> 1300*c(0.4,0.25,0.3,0.05)

[1] 520 325 390  65
```

R calcula de manera simple y directa el estadístico χ^2 que hemos enunciado más arriba utilizando la función `chisq.test()`, a la que se le deben pasar como argumentos las frecuencias observadas y las proporciones esperadas bajo H_0:

```
> tabla = table(playa[sexo == "Hembra"])
> chisq.test(tabla, p = c(0.4, 0.25, 0.3, 0.05))

        Chi-squared test for given probabilities

data:  tabla
X-squared = 635.2, df = 3, p-value < 2.2e-16
```

El valor del estadístico χ^2 es 635.2. El p-valor del contraste (<0.0001) es menor que $\alpha = 0{,}05$, lo que indica que puede rechazarse H_0 (en otras palabras, el valor χ^2 es demasiado grande como para deberse al efecto del azar). Consecuentemente, concluimos que existe evidencia suficiente de que la proporción de nidos por playa ha cambiado desde los años 50.

Respuesta correcta:

Respuesta:		pvalor:	

12. ¿Muestran los datos asociación significativa entre el éxito de emergencia y la presencia de cangrejos? Estima el éxito de emergencia con y sin presencia de cangrejos, acompañando dicha estimación de un intervalo de confianza. Estima también el cociente entre ambas proporciones y construye un intervalo de confianza para dicha estimación.

¿Existe asociación significativa?		pvalor:	

Parámetro	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
π_{CON} =Éxito de emergencia ***con*** cangrejos			
π_{SIN} =Éxito de emergencia ***sin*** cangrejos			
π_{CON}/π_{SIN}			

Solución:

Podemos plantear este problema de dos maneras:

a) Comparando el valor medio del éxito de emergencia entre nidos con cangrejos y nidos sin cangrejos:

$$\begin{cases} H_0: & \mu_{SIN} = \mu_{CON} \\ H_1: & \mu_{SIN} \neq \mu_{CON} \end{cases}$$

Si este valor medio es diferente según que haya o no cangrejos (H_1), ello implicaría que la presencia de cangrejos está asociada con el éxito de emergencia. En caso de aceptar H_0 ello significaría que la presencia o ausencia de cangrejos no influye en el éxito de emergencia.

b) Comparando la proporción de huevos que dan lugar a crías emergidas según que haya o no cangrejos:

$$\begin{cases} H_0: & \pi_{SIN} = \pi_{CON} \\ H_1: & \pi_{SIN} \neq \pi_{CON} \end{cases}$$

Nuevamente, si puede aceptarse que ambas proporciones son iguales (H_0), ello significaría que la probabilidad de que un huevo dé lugar a una cría viva es independiente de que haya o no cangrejos. Por el contrario, rechazar H_0 implica que el éxito de emergencia depende de la presencia o ausencia de cangrejos.

El contraste (a) se resolvería mediante:

```
> table(cangrejos)

cangrejos
 NO  SÍ
365 573

> t.test(exitoEmerg ~ cangrejos)

        Welch Two Sample t-test

data:  exitoEmerg by cangrejos
t = 25.08, df = 735.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
 0.3854 0.4508
sample estimates:
mean in group NO mean in group SÍ
          0.6150           0.1969
```

Hemos construido la primera tabla simplemente para comprobar que el número de nidos es lo suficientemente grande como para que sea válida la aproximación normal, y efectivamente es así (recuérdese que podemos utilizar `t.test()` para comparar las medias de variables no normales siempre que los tamaños muestrales sean grandes, como es el caso, con más de 200 nidos en cada grupo). Los datos muestran que, en presencia de cangrejos, el éxito medio de emergencia es del 19.69 %, frente a un 61.5 % cuando no hay cangrejos en el nido. El p-valor del contraste ($<$0.0001) es menor que $\alpha = 0{,}05$ lo que significa que existe evidencia suficiente para rechazar H_0 y concluir, por tanto, existe una clara asociación entre el éxito de emergencia y la presencia de cangrejos.

El contraste (b) es más artificioso, pues requiere contar el total de huevos depositados en nidos con y sin cangrejos, así como el total de crías emergidas en cada caso:

```
> (nHueCon = sum(Huevos[cangrejos == "SÍ"], na.rm = T))

[1] 47760

> (nHueSin = sum(Huevos[cangrejos == "NO"], na.rm = T))

[1] 30574

> (nEmergCon = sum(crias_Emerg[cangrejos == "SÍ"], na.rm = T))

[1] 9415

> (nEmergSin = sum(crias_Emerg[cangrejos == "NO"], na.rm = T))

[1] 18765

> (p.Con = nEmergCon/nHueCon)

[1] 0.1971

> (p.Sin = nEmergSin/nHueSin)

[1] 0.6138
```

Observemos que la proporción de huevos con éxito reproductivo en cada caso (con y sin cangrejos) es consistente con los valores que hemos estimado antes considerando las medias de éxito de emergencia en todos los nidos. El contraste de proporciones se lleva a cabo mediante `prop.test()`:

```
> prop.test(c(nEmergSin, nEmergCon), c(nHueSin, nHueCon))

        2-sample test for equality of proportions with continuity
        correction

data:  c(nEmergSin, nEmergCon) out of c(nHueSin, nHueCon)
X-squared = 14046, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.4101 0.4232
sample estimates:
prop 1 prop 2
0.6138 0.1971

> ptv = prop.test(c(nEmergSin, nEmergCon), c(nHueSin, nHueCon))$p.value
```

El p-valor de este contraste es también consistente con el anterior (<0.0001) llevándonos por tanto a la misma conclusión.

Por último, para evaluar los intervalos de confianza para el éxito de emergencia con y sin cangrejos, utilizamos `binom.test()`:

```
> binom.test(nEmergSin, nHueSin)$conf.int

[1] 0.6083 0.6192
attr(,"conf.level")
[1] 0.95

> binom.test(nEmergCon, nHueCon)$conf.int

[1] 0.1936 0.2007
attr(,"conf.level")
[1] 0.95
```

y para calcular un intervalo de confianza para el cociente π_{CON}/π_{SIN} (que estimamos mediante `p.Con/p.Sin`=0.3212):

```
> library(PropCIs)
> riskscoreci(nEmergCon, nHueCon, nEmergSin, nHueSin, conf = 0.95)

data:

95 percent confidence interval:
 0.3148 0.3277
```

Vemos así que cuando hay cangrejos en el nido, el éxito de emergencia se reduce hasta ser aproximadamente un 32.12 % del éxito cuando no hay cangrejos.

Respuesta correcta:

¿Existe asociación significativa?		pvalor:	

Parámetro	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
π_{CON} =Éxito de emergencia **con** cangrejos			
π_{SIN} =Éxito de emergencia **sin** cangrejos			
π_{CON}/π_{SIN}			

13. Estima el éxito de emergencia global en toda la isla, proporcionando un intervalo de confianza. ¿Se puede asegurar que el éxito de emergencia global supera el 25 %?

Parámetro	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
π_G =Éxito de emergencia global			

¿Se puede asegurar que $\pi_G > 0{,}25$?		pvalor:	

Solución:

Para estimar el éxito de emergencia global contabilizamos el total de huevos depositados en los nidos, así como el total de crías emergidas de los mismos:

```
> (nHuevos = sum(Huevos, na.rm = T))

[1] 108811

> (nEmerg = sum(crias_Emerg, na.rm = T))
```

```
[1] 28180
```

El éxito global de emergencia será el cociente entre ambas cantidades:

```
> (p.G = nEmerg/nHuevos)

[1] 0.259
```

Para estimar el intervalo de confianza procedemos como en casos anteriores:

```
> binom.test(nEmerg, nHuevos)$conf.int

[1] 0.2564 0.2616
attr(,"conf.level")
[1] 0.95
```

Por último para contrastar si puede asegurarse que $\pi_G > 0{,}25$ utilizamos nuevamente `binom.test()`, especificando `p=0.25` y el sentido de la alternativa:

```
> binom.test(nEmerg, nHuevos, p = 0.25, alternative = "greater")

        Exact binomial test


data:  nEmerg and nHuevos
number of successes = 28180, number of trials = 108811,
p-value = 4.81e-12
alternative hypothesis: true probability of success is greater than 0.25
95 percent confidence interval:
 0.2568 1.0000
sample estimates:
probability of success
                 0.259
```

El p-valor del contraste (<0.0001) es menor que $\alpha = 0{,}05$ lo que significa que existe evidencia suficiente para asegurar que $\pi_G > 0{,}25$; es más, mirando el intervalo de confianza podemos asegurar, con un 95 % de confianza, que π_G es mayor que 0.2568.

Respuesta correcta:

Parámetro	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
π_G =Éxito de emergencia global			

¿Se puede asegurar que $\pi_G > 0{,}25$?		pvalor:	

14. Estima la recta de regresión de ACC frente a LCC para cada sexo. Acompaña la estimación de cada coeficiente mediante su correspondiente intervalo de confianza. ¿Es significativa la pendiente de la regresión en cada caso? Si se elige un macho al azar y resulta que LCC=75 cm ¿qué valor puedes predecir para su ACC? Acompaña dicha predicción de un intervalo de confianza. ¿Cuál es el valor medio de ACC para las hembras cuya LCC es 75 cm? Calcula un intervalo de confianza para dicha predicción.

Machos	Estimación	Intervalo de Confianza		¿Es la pendiente significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
β_0					
β_1					

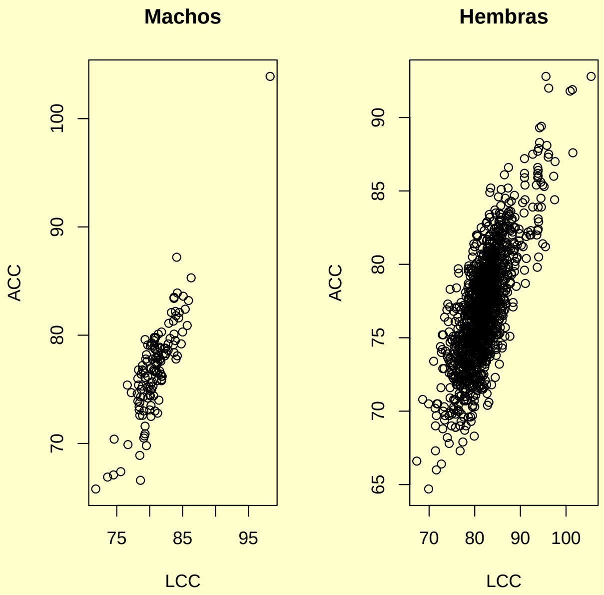
Hembras	Estimación	Intervalo de Confianza		¿Es la pendiente significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
β_0					
β_1					

Valor a estimar	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
ACC de un macho de 75 cm.			
ACC medio de las hembras de 75 cm.			

Solución:

Comenzamos dibujando las nubes de puntos para comprobar visualmente que efectivamente existe asociación lineal entre las variables indicadas:

```
> par(mfrow = c(1, 2))
> with(subset(tortugas, sexo == "Macho"), plot(LCC, ACC,
    main = "Machos"))
> with(subset(tortugas, sexo == "Hembra"), plot(LCC, ACC,
    main = "Hembras"))
```



Se aprecia que efectivamente existe asociación aproximadamente lineal entre las variables (evidentemente no es una asociación perfecta). Para estimar las rectas de regresión para cada sexo utilizamos `lm()`:

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

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

Coefficients:
(Intercept)          LCC
     -32.74         1.36

> rectaH

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

Coefficients:
(Intercept)          LCC
     22.719        0.663
```

En el caso de los machos, la pendiente de la regresión nos informa que por cada cm. de incremento en LCC se producen 1.36 cm. de incremento en ACC; en el caso de las hembras, por cada cm.

que se incrementa la LCC, la ACC se incrementa en 0.66 cm. Las ordenadas representan el valor de ACC cuando LCC es 0; ahora bien, puesto que obviamente la longitud del caparazón de una tortuga no puede ser 0, la ordenada en este caso no tiene más interpretación que la de ser un simple coeficiente de ajuste necesario para que la recta pase por el centro de la nube de puntos.

Los intervalos de confianza para la ordenada y la pendiente se obtienen utilizando `confint()`:

```
> confint(rectaM)

              2.5 %  97.5 %
(Intercept) -43.841 -21.637
LCC           1.219   1.494

> confint(rectaH)

              2.5 %  97.5 %
(Intercept) 20.1634 25.2755
LCC          0.6315  0.6938
```

Estos intervalos, unidos a la estimación de las pendientes, nos permiten apreciar que el incremento en ACC asociado al incremento en LCC en machos prácticamente duplica al que se produce en hembras, lo que constituye probablemente una característica morfológica que permite diferenciar ambos sexos.

Mediante `summary()` obtenemos el análisis estadístico completo de la regresión; en particular, el último p-valor mostrado corresponde al contraste de la regresión:

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$$

Si el p-valor es menor que $\alpha = 0,05$, existe evidencia suficiente para rechazar H_0 y por tanto la pendiente es significativamente distinta de cero; en caso contrario, concluiríamos que la asociación es espuria (el ajuste a la recta se ha producido por azar). A priori, los gráficos anteriores apuntan a que la asociación resultará significativa. Comprobamos:

```
> summary(rectaM)

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

Residuals:
   Min     1Q Median     3Q    Max
 -7.28  -1.34  -0.11   1.76   5.86
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.7388     5.6112   -5.83  4.1e-08 ***
LCC           1.3564     0.0693   19.56  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.28 on 129 degrees of freedom
Multiple R-squared: 0.748,       Adjusted R-squared: 0.746
F-statistic:  383 on 1 and 129 DF,  p-value: <2e-16
```

```
> summary(rectaH)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-7.364 -1.678  0.028  1.787  7.151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.7195     1.3029    17.4   <2e-16 ***
LCC           0.6626     0.0159    41.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.47 on 1298 degrees of freedom
Multiple R-squared: 0.573,       Adjusted R-squared: 0.573
F-statistic: 1.74e+03 on 1 and 1298 DF,  p-value: <2e-16
```

En ambos casos los p-valores son suficientemente pequeños (<0.0001) por lo que concluímos que efectivamente la pendiente es significativa.

Por último, utilizamos la recta ajustada a los machos para predecir el valor de ACC en un macho cuya LCC es 75 cm, así como un intervalo de confianza para dicha predicción:

```
> dataM = data.frame(LCC = 75)
> predict(rectaM, dataM, interval = "prediction", level = 0.95)
```

```
    fit  lwr   upr
1 68.99 64.4 73.59
```

Asimismo, para predecir el valor medio de ACC en hembras cuya LCC es 75 cm. utilizamos:

```
> dataH = data.frame(LCC = 75)
> predict(rectaH, dataH, interval = "confidence", level = 0.95)

    fit   lwr   upr
1 72.42 72.16 72.67
```

Respuesta correcta:

Machos	Estimación	Intervalo de Confianza		¿Es la pendiente significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
β_0					
β_1					

Hembras	Estimación	Intervalo de Confianza		¿Es la pendiente significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
β_0					
β_1					

Valor a estimar	Estimación	Intervalo de Confianza	
		Ext. Inf.	Ext. Sup.
ACC de un macho de 75 cm.			
ACC medio de las hembras de 75 cm.			

15. Estima el coeficiente de correlación lineal entre LCC y el tamaño de la puesta, y entre la distancia a la línea de marea y el tamaño de puesta. Acompaña cada estimador de su correspondiente intervalo de confianza, e indica si la correlación es o no significativa.

Variables	Correlación	Intervalo de Confianza		¿Es la correlación significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
LCC- Puesta					
Dist-Puesta					

Solución:

La función `cor()` calcula la correlación entre variables:

```
> cor(LCC, Huevos, use = "pair")

[1] 0.5012

> cor(distancia, Huevos, use = "pair")

[1] 0.01403
```

(la opción `use="pair"` indica a R que utilice todas las observaciones que no tengan valores perdidos en el par de variables cuya correlación se pretende calcular).

Si además queremos calcular intervalos de confianza y contrastar si la correlación es significativa, utilizamos `cor.test()`:

```
> cor.test(LCC, Huevos)

        Pearson's product-moment correlation

data:  LCC and Huevos
t = 20.86, df = 1298, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4593 0.5408
sample estimates:
   cor
0.5012

> cor.test(distancia, Huevos)

        Pearson's product-moment correlation

data:  distancia and Huevos
t = 0.5055, df = 1298, p-value = 0.6133
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04037  0.06835
sample estimates:
    cor
0.01403
```

En el primer caso el p-valor del contraste ($<$0.0001) es menor que $\alpha = 0{,}05$ lo que significa que existe evidencia suficiente para asegurar que la correlación entre ambas variables es no nula. En el segundo caso ocurre justamente lo contrario, por lo que la correlación no es significativa.

Respuesta correcta:

Variables	Correlación	Intervalo de Confianza		¿Es la correlación significativa?	p-valor
		Ext. Inf.	Ext. Sup.		
LCC- Puesta					
Dist-Puesta					

Dudas y preguntas: Para aclarar todas las dudas y cuestiones que se planteen en el desarrollo de este trabajo, el profesor estará disponible en su horario de tutorías. Pueden realizarse consultas también por correo electrónico, si bien es preferible que se hagan a través del foro dispuesto para ello en el campus virtual, de tal forma que las aclaraciones a las dudas van quedando disponibles en el foro para todo el mundo.

Observación final: los datos suministrados para la realización de este trabajo, si bien se inspiran en datos reales correspondientes a campañas de muestreo que efectivamente se realizaron en las fechas y lugares indicados, son datos simulados. La simulación se ha realizado de tal manera que se asemeje lo más posible al comportamiento general observado en la realidad. Se han modificado, no obstante, muchas características: proporción de ejemplares por playas, relaciones morfológicas, tasas de éxito de emergencia, etc. De esta forma el archivo de datos que recibe cada alumno para la realización del trabajo es diferente del resto de los archivos, por lo que las relaciones y características destacables que se puedan encontrar pueden diferir notablemente de unos a otros.