

1. (3 ptos.) En el ensayo clínico de un test diagnóstico para una enfermedad, se anotó para cada individuo su estado (enfermo E o sano S) y el resultado (+ o -) en la prueba. Datos en <https://goo.gl/JPYqg4>

Descarga de datos y calcula de la tabla de contingencia

```
direccion = "https://goo.gl/JPYqg4"
datos = read.table(file = direccion, header = T, sep = ";")
datos = datos[complete.cases(datos), ]
(tabla = table(datos))

##      estado
## test   E   S
##    -  59 460
##    + 436  35
```

- (a) (1.5 ptos) Calcula la sensibilidad, especificidad y los valores predictivo positivo y negativo del test.

Dedicamos parte de la práctica 5 a resolver un ejercicio que incluía estas preguntas.

En términos de probabilidades, lo que se pide es

- Sensibilidad del test: $P(+|E)$

```
tabla[2,1]/sum(tabla[,1])
## [1] 0.8808081
```

- Especificidad del test: $P(-|S)$

```
## [1] 0.9292929
```

- Valor predictivo positivo del test: $P(E|+)$

```
## [1] 0.92569
```

- Valor predictivo negativo del test: $P(S|-)$

```
## [1] 0.8863198
```

- (b) (0.5 ptos) Se quiere tomar una muestra de los individuos que han participado en el ensayo, ¿cómo se debe proceder para que todos tengan la misma probabilidad de ser elegidos?

Dedicamos los ejercicios del 6 al 10 de la práctica 4 a este asunto.

Hay que hacer extracciones con reemplazamiento. De esta manera, las extracciones son independientes, porque cada una no altera el número de casos posibles y favorables.

- (c) (1 pto) Una vez aclarado el método, se toma una muestra independiente de 15 individuos, calcula la probabilidad de que al menos 13

sean bien clasificado sabiendo que al menos clasificó bien a 5. La variable de interés es $X = \text{"número de individuos bien clasificados de los 15 seleccionados en la muestra"}$. Como las extracciones son independientes y cada individuo sólo puede estar bien o mal clasificado. En definitiva, $X \sim B(15, p)$ y hay que estimar p . Un individuo está bien clasificado tanto cuando está enfermo y da positivo como cuando está sano y da negativo en el test. Por tanto, calculando los casos favorables entre los casos posibles se tiene

```
(p = (tabla[1,2]+tabla[2,1])/sum(tabla))
## [1] 0.9050505
```

El cálculo que se puede es $P(X \geq 13 | X \geq 5) = \frac{P(X \geq 13)}{P(X \geq 5)}$, es decir

```
sum(dbinom(13:15, size = 15, prob = p))/
  sum(dbinom(5:15, size = 15, prob = p))
## [1] 0.8350676
```

2. (2 ptos.) Un método indirecto para estimar la cantidad de cierto contaminante en una laguna consiste en medir la presencia/ausencia de un organismo sensible al contaminante (biosensor). El procedimiento consiste en tomar 10 muestras independientes y si el biosensor está presente:
 - (a) (1 pto) Si el biosensor está presente en 7 o más muestras, la concentración de contaminante (mg/l) se distribuye conforme a $f(x) = e^{-x/10}/10$ si $x \geq 0$, $f(x) = 0$ si $x < 0$
 - (b) (1 pto) Si el biosensor está presente en entre 3 y 6 muestras (ambas incluidas), se distribuye de forma uniforme en el intervalo $[3, 23]$.
 - (c) Si el biosensor está presente en 2 o menos muestras se distribuye conforme a una normal $N(17, 4)$

La probabilidad de detectar el biosensor es 0.3 y la contaminación es crítica a partir de 19mg/l. Se hace el experimento una vez: calcula

- (a) La probabilidad de que el nivel de contaminante NO sea crítico. Llamaremos $W = \text{"cantidad de contaminante"}$, y $X = \text{"número de muestras de agua en las que se detecta el biosensor, de entre las 10"}$ se pide $P(W < 19)$. Esto puede suceder en tres escenarios diferentes: cuando se ha detectado el biosensor en 8 o más muestras, en entre 3 y 7 (ambos incluidos) o en 2 o menos muestras. Es decir, con la regla de la probabilidad total esto es

$$\begin{aligned} P(W < 19) &= P(X \leq 2) * P(W < 19 | X \leq 2) \\ &\quad + P(3 \leq X \leq 7) * P(W < 19 | 3 \leq X \leq 7) \\ &\quad + P(8 \leq X) * P(W < 19 | 8 \leq X) \end{aligned}$$

Es decir, cada sumando es

```
sum1 = sum(dbinom(7:10, size = 10, prob = 0.3))*pexp(19, rate = 1/10)
sum2 = sum(dbinom(3:6, size = 10, prob = 0.3))*punif(19, 3, 23)
sum3 = pbinom(2, size = 10, prob = 0.3)*pnorm(19, 17, 4)
```

la probabilidad de la exponencial se podía calcular con wolframalpha, claro. La probabilidad pedida es

```
(noContaminado = sum1 + sum2 + sum3)

## [1] 0.7589879
```

- (b) Si el nivel de contaminante es crítico, ¿cuál es la probabilidad de que se detectara el biosensor en 2 o menos muestras?
Esto es una aplicación del teorema de Bayes:

$$P(X \leq 2 | \text{contaminado}) = \frac{P(X \leq 2 \cap \text{contaminado})}{P(\text{contaminado})}$$

se puede reaprovechar los cálculos del apartado anterior

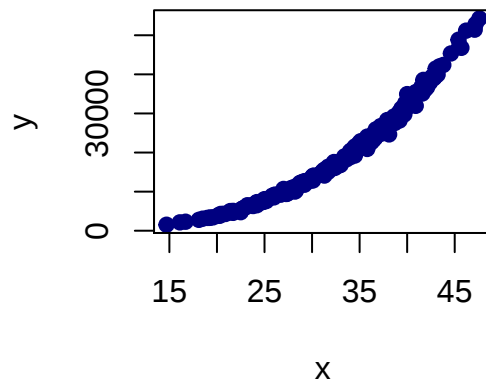
```
numerador = pbinom(2, size = 10, prob = 0.3)*(1-pnorm(19, mean = 17, sd = 4))
denominador = 1-noContaminado
numerador/denominador

## [1] 0.4900287
```

3. (Ejercicio opcional, 1 pto.). En un pendrive olvidado en un cajón ha aparecido una nueva versión del fichero `crabs` con una variable adicional llamada `CV`, ahora disponible en <https://goo.gl/iPT3Jd> Se sospecha que la nueva variable `CV` puede estar relacionada con la longitud del caparazón `CL` de cada individuo:

- (a) Representa la nueva variable `CV` en función de `CL`.

```
direccion3 = "https://goo.gl/iPT3Jd"
newcraws = read.table(file = direccion3, header = T, sep = ";")
x = newcraws$CL; y = newcraws$CV
plot(x, y, pch = 19, col = "navy")
```



- (b) Decide qué relación (lineal, logarítmica, exponencial, potencial) describe mejor dicha relación.

A simple vista se pueden descartar la lineal y la logarítmica. La exponencial y la potencial tienen por ecuación, respectivamente

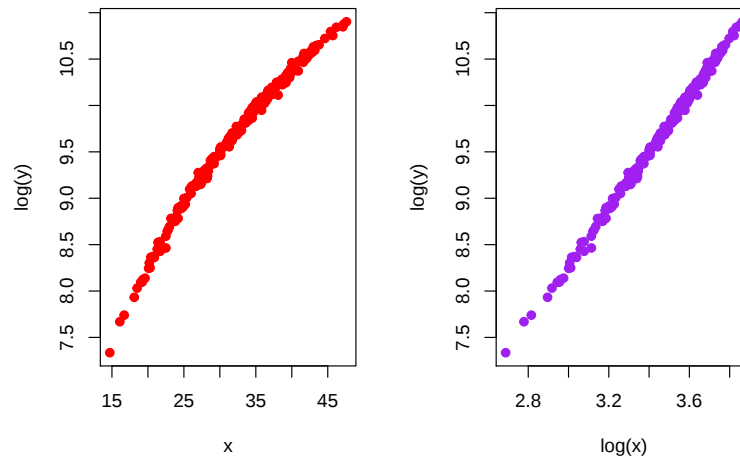
$$y = b_0 e^{b_1 x}, \quad y = b_0 x^{b_1}$$

Para transformar esas expresiones en una recta hay que tomar logaritmos (usaremos el neperiano). En el primer caso queda

$$\log(y) = \log(b_0) + b_1 x, \quad \log(y) = \log(b_0) + b_1 \log(x)$$

lo que sugiere representar x frente a $\log(y)$ y $\log(x)$ frente a $\log(y)$:

```
par(mfrow= c(1,2))
plot(x, log(y), pch = 19, col = "red")
plot(log(x), log(y), pch = 19, col = "purple")
```



```
par(mfrow= c(1,1))
```

El modelo potencial parece mejor; podemos calcular los coeficientes de correlación para corroborar esa impresión visual

```
cor(x, log(y))
## [1] 0.9884834
cor(log(x), log(y))
## [1] 0.9986211
```

apenas hay diferencia, pero es más alto el del modelo potencial.

- (c) Escribe la correspondiente expresión (fórmula). Contruir el modelo para obtener los coeficientes

```
lmPot = lm(log(y) ~ log(x))
lmPot$coefficients

## (Intercept)      log(x)
## -0.8935435    3.0544396

(b0 = exp(unname(lmPot$coefficients[1])))
## [1] 0.4092032

(b1 = unname(lmPot$coefficients[2]))
## [1] 3.05444
```

y la fórmula es

$$newcrabs\$CV = 0.409 * newcrabs\$CL^{3.05}$$

- (d) Se conjetura que la nueva variable es volumen del cangrejo; ¿te parece plausible?

Sí, puesto que el volumen se expresa en unidades de longitud elevado a 3. Es razonable que al aumentar la longitud del caparazón el volumen aumente según su cubo, el volumen es tridimensional y la longitud unidimensional.