Examen final de ESTADÍSTICA APLICADA (20 de mayo de 2019)
Primer curso del grado en Bioquímica
SOLUCIONES

---

1. e)

$$X = \begin{cases} 1 & \text{si una bellota está contaminada con } \textit{E. coli} \\ 0 & \text{si no lo está} \end{cases} \sim \text{Bernoulli}(p_1)$$

$$Y = \begin{cases} 1 & \text{si una tableta está contaminada con } \textit{E. coli} \\ 0 & \text{si no lo está} \end{cases} \sim \text{Bernoulli}(p_2)$$

$$\hat{p}_1 = \frac{40}{43} = 0.9302326 \qquad \hat{p}_2 = \frac{5}{17} = 0.294176$$

a) Un intervalo de confianza al 95 % para la proporción p_1 es

$$\text{IC}_{95\,\%}(p_1) = \left(0.9302 \mp 1.96\sqrt{\frac{0.9302(1-0.9302)}{43}}\right) = (0.9302 \mp 0.0761).$$

b) Al nivel de significación $\alpha = 0.05$, queremos contrastar

$$\begin{aligned} H_0: &\quad p_1 \leq 0.75 \\ H_1: &\quad p_1 > 0.75. \end{aligned}$$

La región de rechazo es $R = \{z > z_{0.05} = 1.64\}$, donde el estadístico del contraste es

$$z = \frac{0.9302 - 0.75}{\sqrt{\frac{0.75\ 0.25}{43}}} = 2.729.$$

Por tanto, hay suficiente evidencia muestral para rechazar H_0 a ese nivel.

c) El código resuelve el contraste

$$\begin{aligned} H_0: &\quad p_1 = 0.75 \\ H_1: &\quad p_1 \neq 0.75. \end{aligned}$$

La salida de R proporciona:
`X-squared`, que es el estadístico del contraste, que se calcula como z^2, aunque en este caso incluye una corrección por continuidad (de ahí que no coincida con el z^2 del z calculado en (b)).
El p-valor del contraste (0.01067). Luego rechazamos H_0 para todos los α's mayores que el p-valor, por ejemplo, $\alpha = 0.05$ y $\alpha = 0.1$, pero no para $\alpha = 0.01$. Por tanto, es una situación en la que no está clara la conclusión acerca de si rechazar o aceptar H_0.
$\text{IC}_{90\,\%}(p_1) = (0.8228321, 0.9783592)$
$\hat{p}_1 = \bar{x} = 0.9302326$

d) La afirmación (*) no se ha obtenido por un procedimiento correcto. El contraste para comparar dos proporciones

$$\begin{aligned} H_0: &\quad p_1 \leq p_2 \\ H_1: &\quad p_1 > p_2 \end{aligned}$$

tiene una región de rechazo construida utilizando el TCL y sólo es válida para tamaños muestrales suficientemente grandes en ambas poblaciones. En el caso de X el tamaño muestral es 43, que es suficientemente grande, pero en el caso de Y el tamaño muestral es 17, que no es suficientemente grande para que el procedimiento sea válido.

2. a) Las variables X e Y son dependientes porque se han medido en las mismas focas, dando lugar a $n = 10$ pares de datos emparejados (x_i, y_i), $i = 1, \ldots, 10$. Suponemos que $D = X - Y$ sigue una distribución N(μ,σ) con μ y σ desconocidos. Los valores observados de D aparecen en la Tabla 1.

Foca	No alimentaria	Alimentaria	d_i	Foca	No alimentaria	Alimentaria	d_i
1	42.2	71.0	-28.8	6	82.0	112.8	-30.8
2	51.7	77.3	-25.6	7	81.3	121.2	-39.9
3	59.8	82.6	-22.8	8	81.3	126.4	-45.1
4	66.5	96.1	-29.6	9	96.0	127.5	-31.5
5	81.9	106.6	-24.7	10	104.1	143.1	-39.0

Tabla 1

La media y desviación típica muestrales de D son respectivamente $\bar{d} = -31.78$ y $s_d = 7.2962$.
Si denotamos $\mu_1 = E(X)$ y $\mu_2 = E(Y)$, un intervalo de confianza al 99 % para la diferencia de los costes metabolicos medios entre una inmersión no alimentaria y otra alimentaria es

$$\mathrm{IC}_{99\,\%}(\mu_1 - \mu_2) = \mathrm{IC}_{99\,\%}(\mu) = \left(-31.78 \mp 3.25 \frac{7.2962}{\sqrt{10}}\right) = (-31.78 \mp 7.50).$$

b) (1 punto) A un nivel de significación del 10 %, ¿hay evidencia a favor de la hipótesis de que la inmersión alimentaria causa en promedio un mayor coste metabólico que una no alimentaria? ¿Y a niveles de significación del 5 % y del 1 %? ¿Qué puedes decir del p-valor del contraste?
Utilizamos las mismas suposiciones que en (a). Queremos hacer el contraste

$$\begin{array}{ll} H_0: & \mu_1 \geq \mu_2 \\ H_1: & \mu_1 < \mu_2 \end{array} \quad \text{o lo que es equivalente} \quad \begin{array}{ll} H_0: & \mu \geq 0 \\ H_1: & \mu < 0 \end{array}$$

La región de rechazo es $R = \{t < t_{9;1-\alpha}\}$, donde el estadístico del contraste es

$$t = \frac{-31.78}{7.2962/\sqrt{10}} = -13.77.$$

Como $t_{9;0.99} = -t_{9;0.01} = -2.82$, rechazamos H_0 a nivel $\alpha = 0.01$ y, por tanto, también a los niveles mayores que 0.01, como $\alpha = 0.05$ y $\alpha = 0.1$. El p-valor es menor que 0.01 pues es el ínfimo nivel de significación para el que rechazamos H_0. Como $t_{9;0.0005} = 4.781$, también rechazamos H_0 al nivel $\alpha = 0.0005$ y concluimos que el p-valor es menor que 0.0005.

c) El coeficiente de correlación entre los costes metabólicos en una inmersión alimentaria y una no alimentaria es

$$r = \frac{\mathrm{cov}_{x,y}}{s_x s_y} = \frac{451.6824}{\sqrt{376.0884\ 580.5116}} = 0.97.$$

d) La pendiente de la recta de regresión es

$$b = \frac{\mathrm{cov}_{x,y}}{s_x^2} = \frac{451.6824}{376.0884} = 1.20.$$

La ordenada en el origen es

$$a = \bar{y} - b\bar{x} = 106.46 - 1.20\ 74.68 = 16.84.$$

La recta de regresión pedida es, pues, $y = 16.84 + 1.2x$.

e) El gráfico pedido aparece en la Figura 1 y muestra un buen ajuste de los datos a la recta de regresión. Esto es coherente con la correlación obtenida en (c), que tiene un valor muy próximo a 1, lo cual es indicativo de un alto grado de relación lineal entre las dos variables.

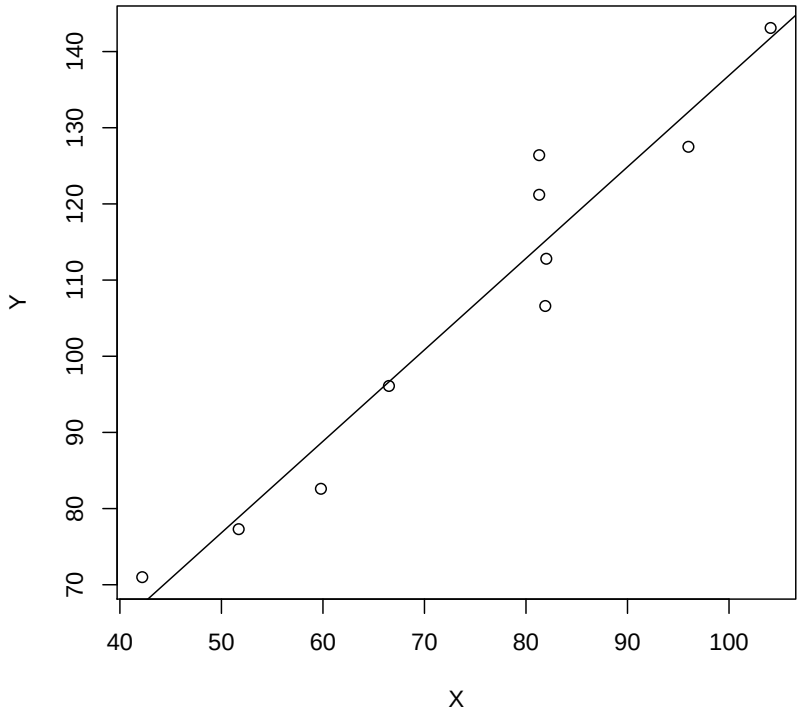

Figura 1

f) En los diagramas de caja observamos que el primer cuartil del coste metabólico en una inmersión alimentaria parece superior al tercer cuartil del coste metabólico en una inmersión no alimentaria. Por tanto, un 75 % de los datos de Y están por encima de un 75 % de los datos de X. Además la mediana de Y es mayor que el máximo de X (dado por el límite superior del diagrama de caja), luego la mitad de los datos de Y es mayor que todos los datos de X. En consecuencia parece razonable que el valor esperado del coste metabólico en una inmersión alimentaria sea mayor que aquél en una inmersión no alimentaria, que es justo la hipótesis alternativa que habíamos aceptado en (b).

3. $X \sim N(\mu = 5.6, \sigma = 2.3) \quad Z \sim N(0,1)$

a) La probabilidad de que X esté entre 3 y 8 es

$$P\{3 < X < 8\} = P\{-1.13 < Z < 1.04\} = 1 - 0.1492 - 0.1292 = 0.7216.$$

b) La media $\bar{X} = (X_1 + X_2 + X_3)/3$ de las tres medidas sigue una distribución N($\mu = 5.6$, $2.3/\sqrt{3} = 1.33$). La probabilidad de que $\bar{X}$ esté entre 3 y 8 es

$$P\{3 < \bar{X} < 8\} = P\{-1.95 < Z < 1.80\} = 0.9385.$$

c)

$$0.25 = P\{X < Q_1\} = P\left\{Z < \frac{Q_1 - 5.6}{2.3}\right\} = P\left\{Z > \frac{5.6 - Q_1}{2.3}\right\}$$

Por tanto,

$$\frac{5.6 - Q_1}{2.3} = 0.675 \Rightarrow Q_1 = 4.0475$$

d) La variable $Y =$ "número de esas diez medidas que están entre 3 y 8" sigue una distribución binomial B(10,0.7216).

$$P\{Y \leq 7\} = 1 - (P\{Y = 8\} + P\{Y = 9\} + P\{Y = 10\}) = 1 - 0.2564 - 0.1477 - 0.0383 = 0.5576.$$