Problema 1

Examen Final

29 de mayo de 2023

Se está realizado un estudio sobre la distribución de la concentración X de trombocitos (en miles por milímetro cúbico) en pacientes de cierta enfermedad. Se ha obtenido una muestra de tales pacientes con el siguiente resultado: `c(88.4, 120.4, 99.2, 96.4, 110.1, 110.2, 86.2, 94.7, 108.9, 111.5, 105.6, 105.3, 106.7, 107.9, 123.4, 121.5, 85.6, 115.2, 124.9, 94, 111.4, 100.8, 106.4, 104.7, 111.8, 139.6, 130, 145.5, 128, 137.8, 119, 118.7, 108.5, 122.7, 112.9, 116.8, 144.5, 110.1, 129, 110.4, 104.1, 129.3, 98.7, 118.5, 111, 103.4, 92.4, 95.2, 105.7, 130, 101.5, 100.6, 133.9, 139.7, 132.2, 135.5, 113.6, 101.2, 96.4, 104)`

Para los siguientes apartados, supón que $X \hookrightarrow \mathcal{U}\left(\frac{1}{2}\theta;\theta\right)$.

1. Halla un estadígrafo mínimamente suficiente para θ.

   $$X \hookrightarrow \mathcal{U}\left(\frac{1}{2}\theta;\theta\right) \implies f(x) = \frac{1}{\theta - \frac{1}{2}\theta} \cdot \mathbb{1}\left(\frac{1}{2}\theta \leqslant x \leqslant \theta\right) = \frac{2}{\theta} \cdot \mathbb{1}\left(\frac{1}{2}\theta \leqslant x \leqslant \theta\right) \implies$$
   $$\mathcal{L}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i) = \frac{2^n}{\theta^n} \cdot \mathbb{1}\left(\frac{1}{2}\theta \leqslant x_{(1)}\right) \cdot \mathbb{1}\left(x_{(n)} \leqslant \theta\right) =: g(t,\theta) \text{ con } t := (x_{(1)}, x_{(n)})$$
   $$\implies T := T(\vec{X}) := (X_{(1)}, X_{(n)}) \text{ es suficiente.}$$

   $\vec{x} \neq \vec{y}$ muestras aleatorias simples de tamaño $n \implies \dfrac{\mathcal{L}(\vec{x})}{\mathcal{L}(\vec{y})} = \dfrac{\mathbb{1}\left(\frac{1}{2}\theta \leqslant x_{(1)}\right) \cdot \mathbb{1}\left(x_{(n)} \leqslant \theta\right)}{\mathbb{1}\left(\frac{1}{2}\theta \leqslant y_{(1)}\right) \cdot \mathbb{1}\left(y_{(n)} \leqslant \theta\right)}$ independiente de θ sii $T(\vec{x}) = T(\vec{y})$ pues entonces las indicatrices se cancelan; por ejemplo,

   - $\vec{x} := (10,11,15) \neq (10,13,15) =: \vec{y} \implies T(\vec{x}) = T(\vec{y}) \implies \dfrac{\mathcal{L}(\vec{x})}{\mathcal{L}(\vec{y})}$ no depende de θ
   - $\vec{x} := (11,15) \neq (13,15) =: \vec{y} \implies T(\vec{x}) \neq T(\vec{y}) \implies \dfrac{\mathcal{L}(\vec{x})}{\mathcal{L}(\vec{y})}$ depende de θ, pues $11 < \theta < 13 \implies \dfrac{\mathcal{L}(\vec{x})}{\mathcal{L}(\vec{y})} = 0$ pero $13 < \theta < 15 \implies \dfrac{\mathcal{L}(\vec{x})}{\mathcal{L}(\vec{y})} = 1$

   Por tanto, T es mínimamente suficiente. Su valor (véase código fuente R al final) para la realización muestral del enunciado es $t = (85.6, 145.5)$

2. Halla la estimación de θ por el método de los momentos.

   $$\mathcal{E}[X] = \frac{\frac{\theta}{2} + \theta}{2} = \frac{3}{4}\theta \implies \hat{\theta}_{\text{MM}} = \frac{4}{3}\bar{X} \implies \hat{\theta}_{\text{MM}}(\vec{x}) = 150.48$$

3. Halla la estimación máximo-verosímil de θ.

   $$\mathcal{L}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i) = \frac{2^n}{\theta^n} \cdot \mathbb{1}\left(\frac{1}{2}\theta \leqslant x_{(1)}\right) \cdot \mathbb{1}\left(x_{(n)} \leqslant \theta\right) \text{ decreciente en } \theta \text{ para}$$
   $x_{(n)} \leqslant \theta \leqslant 2x_{(1)} \implies$ se maximiza en $\hat{\theta}_{\text{MV}} = X_{(n)} \implies \hat{\theta}_{\text{MV}}(\vec{x}) = 145.5$

4. Calcula un intervalo de confianza para θ al $95\,\%$.

Podría usarse

- algún método bútstrap;
- un pivote basado en el método delta aplicado al estimador de los momentos;

pero el mejor pivote suele encontrarse con el estimador máximo–verosímil, en este caso, $X_{(n)}$.

$X \hookrightarrow \mathcal{U}\left(\frac{\theta}{2};\theta\right) \implies Y := 2\left(\frac{X}{\theta} - \frac{1}{2}\right) = \frac{2X}{\theta} - 1 \implies (*)$

$(*) \implies Y \hookrightarrow \mathcal{U}(0;1) \implies F_Y(y) = y$ para $0 < y < 1 \implies (**)$

$(*)$ trasformación estrictamente creciente $\implies Y_{(n)} = \frac{2X_{(n)}}{\theta} - 1 \implies (**)$

$(**) \implies$ para $0 < y < 1$ $F_{Y_{(n)}}(y) = y^n \implies Y_{(n)}$ pivote. Sean a y b cuantiles de $Y_{(n)}$ tales que $1 - \alpha = \Pr\left[a \leqslant Y_{(n)} \leqslant b\right] = \Pr\left[a \leqslant \frac{2X_{(n)}}{\theta} - 1 \leqslant b\right] =$

$\Pr\left[a+1 \leqslant \frac{2X_{(n)}}{\theta} \leqslant b+1\right] = \Pr\left[\frac{1}{a+1} \geqslant \frac{\theta}{2X_{(n)}} \geqslant \frac{1}{b+1}\right] = \Pr\left[\frac{2X_{(n)}}{a+1} \geqslant \theta \geqslant \frac{2X_{(n)}}{b+1}\right]$

$\implies$ I.C.$(\theta) = \left[\frac{2X_{(n)}}{b+1}, \frac{2X_{(n)}}{a+1}\right] = 2X_{(n)}\left[\frac{1}{b+1}, \frac{1}{a+1}\right]$.

Para obtener el intervalo de longitud mínima se tendrá en cuenta que $1 - \alpha = \Pr\left[a \leqslant Y_{(n)} \leqslant b\right] = F_{Y_{(n)}}(b) - F_{Y_{(n)}}(a) = b^n - a^n \implies b = \sqrt[n]{1-\alpha_b}$, $a = \sqrt[n]{\alpha_a}$ con $\alpha_a + \alpha_b = \alpha$. La amplitud A del intervalo $[a, b]$

$$\begin{aligned} A(\alpha_a) &\propto \frac{1}{a+1} - \frac{1}{b+1} = \frac{1}{\sqrt[n]{\alpha_a}+1} - \frac{1}{\sqrt[n]{1-\alpha+\alpha_a}+1} \\ &= \frac{1}{\alpha_a^{\frac{1}{n}}+1} - \frac{1}{(1-\alpha+\alpha_a)^{\frac{1}{n}}+1} \end{aligned}$$

es decreciente[1] pues (denotando la variable por x en vez de por α_a)

$$\begin{aligned} 0 \quad &\geqslant \quad A'(x) = \left\{\frac{1}{x^{\frac{1}{n}}+1} - \frac{1}{(1-\alpha+x)^{\frac{1}{n}}+1}\right\}' \\ &= \quad \frac{-\frac{1}{n}x^{\frac{1}{n}-1}}{(x^{\frac{1}{n}}+1)^2} + \frac{\frac{1}{n}(1-\alpha+x)^{\frac{1}{n}-1}}{[(1-\alpha+x)^{\frac{1}{n}}+1]^2} \\ &= \quad \frac{1}{n}\left\{\frac{-x^{\frac{1}{n}-1}}{(x^{\frac{1}{n}}+1)^2} + \frac{(1-\alpha+x)^{\frac{1}{n}-1}}{[(1-\alpha+x)^{\frac{1}{n}}+1]^2}\right\} \\ &= \quad \frac{1}{n}\left\{\frac{-x^{\frac{1-n}{n}}}{(x^{\frac{1}{n}}+1)^2} + \frac{(1-\alpha+x)^{\frac{1-n}{n}}}{[(1-\alpha+x)^{\frac{1}{n}}+1]^2}\right\} \\ &= \quad \frac{1}{n}\left\{\frac{-1}{x^{\frac{n-1}{n}}(x^{\frac{1}{n}}+1)^2} + \frac{1}{(1-\alpha+x)^{\frac{n-1}{n}}[(1-\alpha+x)^{\frac{1}{n}}+1]^2}\right\} \\ \iff \quad & x^{\frac{n-1}{n}}(x^{\frac{1}{n}}+1)^2 \leqslant (1-\alpha+x)^{\frac{n-1}{n}}[(1-\alpha+x)^{\frac{1}{n}}+1]^2 \\ \iff \quad & \left(\frac{x}{1-\alpha+x}\right)^{\frac{n-1}{n}} \leqslant 1 \leqslant \left(\frac{(1-\alpha+x)^{\frac{1}{n}}+1}{x^{\frac{1}{n}}+1}\right)^2 \end{aligned}$$

[1] ...lo que corresponde con la intuición de que también la densidad de $Y_{(n)}$, $f = n \cdot y^{n-1} \cdot \mathbb{1}(0 < y < 1)$, es creciente en su soporte.

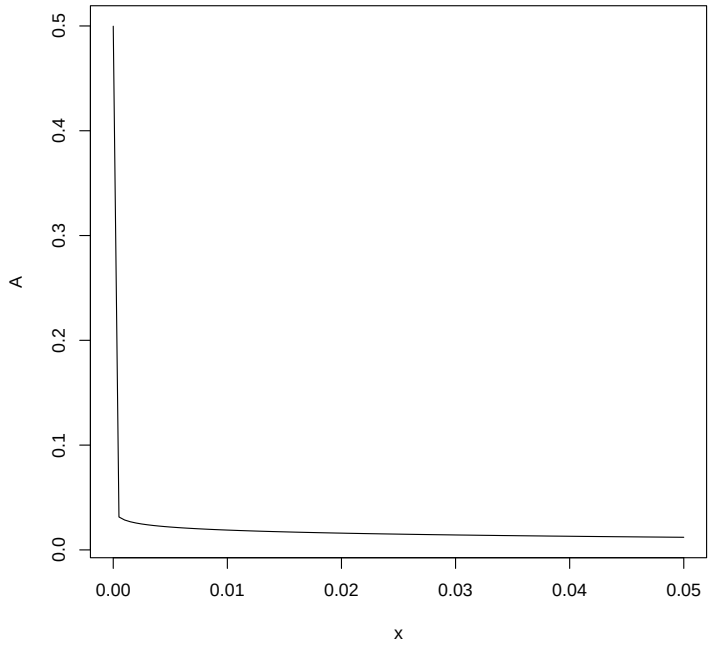
pues

$$\left(\frac{x}{1-\alpha+x}\right)^{\frac{n-1}{n}} \leqslant 1 \iff \frac{x}{1-\alpha+x} \leqslant 1$$
$$\iff x \leqslant 1-\alpha+x$$
$$\iff 0 \leqslant 1-\alpha$$

lo que es obvio pues $0 < x = \alpha_a < \alpha < 1$; por otro lado,

$$\left(\frac{(1-\alpha+x)^{\frac{1}{n}}+1}{x^{\frac{1}{n}}+1}\right)^2 \geqslant 1 \iff \frac{(1-\alpha+x)^{\frac{1}{n}}+1}{x^{\frac{1}{n}}+1} \geqslant 1$$
$$\iff (1-\alpha+x)^{\frac{1}{n}}+1 \geqslant x^{\frac{1}{n}}+1$$
$$\iff (1-\alpha+x)^{\frac{1}{n}} \geqslant x^{\frac{1}{n}}$$
$$\iff 1-\alpha+x \geqslant x$$
$$\iff 1-\alpha \geqslant 0$$

Por tanto, $A' \leqslant 0$ y la amplitud es decreciente:



Por tanto, el mejor intervalo se encuentra con $\alpha_a = \alpha$ y $\alpha_b = 0$, luego $b = 1$, $a = \sqrt[60]{0.05} = 0.9512971$ e I.C. = $[145.5000, 149.1316]$.

Otra alternativa es aprovechar que el tamaño de muestra $n = 60$ es bastante grande y aplicar el TCL:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0;1)$$

$$\frac{\bar X-\mu}{\sigma/\sqrt n}=\frac{\bar X-\frac34\theta}{\frac{1}{\sqrt{48}}\theta/\sqrt n}=\sqrt{48n}\left(\frac{\bar X}{\theta}-\frac34\right)$$

$$1-\alpha=\Pr\left[z_{\alpha_1}\leqslant\sqrt{48n}\left(\frac{\bar X}{\theta}-\frac34\right)\leqslant z_{1-\alpha_2}\right]=\Pr\left[\frac{\bar X}{\frac34+\frac{z_{1-\alpha_2}}{\sqrt{48n}}}\leqslant\theta\leqslant\frac{\bar X}{\frac34+\frac{z_{\alpha_1}}{\sqrt{48n}}}\right]$$

Con $\alpha_1=\alpha_2=\dfrac12\alpha$, I.C. = $[143.4925, 158.1828]$, que incluye valores imposibles, aquéllos menores que $x_{(n)}=145.5$.

5. ¿Se puede considerar que la distribución es uniforme $\mathcal{U}(a;b)$?

Por ser la uniforme continua usaremos Kolgogórov–Smirnov y, por ser la hipótesis nula compuesta, el método de Lilliefors, que da una aproximación al p–valor similar a 0.123616 :

```
Dn <- function (x) ks.test (x, punif, min(x), max(x)) $ statistic
suppressWarnings (mean (replicate (1e6, Dn(runif(n))) >= Dn(x)))
```

Es factible el método de Lilliefors porque la $\mathcal{U}(a;b)$ depende sólo de parámetros asociados a localización y escala (no cambia la forma):

Sean $\hat b=\max X_i$, $\hat a=\hat b/2$, $Y_i:=(X_i-a)/(b-a)\hookrightarrow\mathcal{U}(0;1)$; entonces:

$$\begin{aligned}
D_n &= \sup_{x\in\mathbb{R}}\left|F_n(x)-F_{\mathcal{U}(\hat a;\hat b)}(x)\right|=\sup_{\hat a<x<\hat b}\left|\frac{\sum\mathbb{1}(X_i\leqslant x)}{n}-F_{\mathcal{U}(0;1)}\left(\frac{x-\hat a}{\hat b-\hat a}\right)\right|\\
&= \sup_{\hat a<x<\hat b}\left|\frac{\sum\mathbb{1}\left(\frac{X_i-a}{b-a}\leqslant\frac{x-a}{b-a}\right)}{n}-F_{\mathcal{U}(0;1)}\left(\frac{\frac{x-a}{b-a}-\frac{\hat a-a}{b-a}}{\frac{\hat b-a}{b-a}-\frac{\hat a-a}{b-a}}\right)\right|\\
&= \sup_{\min Y_i<y<\max Y_i}\left|\frac{\sum\mathbb{1}(Y_i\leqslant y)}{n}-F_{\mathcal{U}(0;1)}\left(\frac{y-\min Y_i}{\max Y_i-\min Y_i}\right)\right|
\end{aligned}$$

luego D_n es de libre distribución.

6. Suponiendo que sigue una distribución $\mathcal{U}(a;b)$ contrasta $H_0\colon a=80, b=150$ frente a $H_1\colon a=80, b<150$.

La familia $\mathcal{U}(80;b)$ tiene razón de verosimilitud monótona en $T=X_{(n)}=\max X_i$ pues, dados $b_1<b_2$,

$$\frac{L(b_2)}{L(b_1)}=\begin{cases}0 & b_1<x_{(n)}<b_2\\ \infty & b_2<x_{(n)}\end{cases}$$

Por tanto, la región crítica para el contraste del enunciado tiene la forma $\{\vec x\mid x_{(n)}<c\}\ \exists\, c$. Bajo H_0 la ojiva de $X_{(n)}$ es

$$F_{X_{(n)}}(x)=F_X(x)^n=\left(\frac{x-80}{150-80}\right)^n\qquad\text{para } 80<x<150$$

luego el P –valor para la muestra del enunciado es

$$\Pr\left[X_{(n)}<145.5\right]=F_{X_{(n)}}(x)=\left(\frac{145.5-80}{150-80}\right)^{60}=0.01856075<0.05=\alpha$$

y por tanto se rechaza la hipótesis nula: hay evidencias de que el verdadero valor de b es menor que 150 .

Una aproximación al P –valor puede obtenerse cómodamente aplicando Montecarlo al estadígrafo de la razón de verosimilitudes:

```
RV <- function (x) prod(dunif(x,80,150)) / prod(dunif(x,80,max(x)))
mean (replicate (1e6, RV(runif(n,80,150))) <= RV(x)) # 0.018487
```

7. Si la concentración es menor que 90, el paciente está en riesgo de trombocitopenia. Sea p la proporción poblacional de tales pacientes. Contrasta $H_0\colon p = 0.13$ frente a $H_1\colon p > 0.13$.

   La familia $Y := \mathbb{1}(X < 90) \hookrightarrow \mathcal{B}(n;p)$ tiene razón de verosimilitud monótona en p: si $p_1 < p_2$

$$\frac{\mathcal{L}(p_2)}{\mathcal{L}(p_1)} = \left(\frac{p_2}{p_1}\right)^{\sum y_i} \cdot \left(\frac{1-p_2}{1-p_1}\right)^{n-\sum y_i} = \left(\frac{p_2}{p_1}\right)^{\sum y_i} \cdot \left(\frac{1-p_2}{1-p_1}\right)^{n} \cdot \left(\frac{1-p_1}{1-p_2}\right)^{\sum y_i} \propto$$

$$\propto \left(\frac{p_2}{p_1}\right)^{\sum y_i} \cdot \left(\frac{1-p_1}{1-p_2}\right)^{\sum y_i} = \left(\frac{p_2}{p_1} \cdot \frac{1-p_1}{1-p_2}\right)^{\sum y_i}$$

   creciente en $\sum y_i$ pues $\dfrac{p_2}{p_1}\dfrac{1-p_1}{1-p_2} > 1$.

   Por consiguiente, el contraste uniformemente más potente tiene una región crítica de la forma $\left[\sum Y_i > c\right]$, $\exists\, c$. En la muestra del enunciado, $\sum y_i = 3$ luego el P–valor será

$$\Pr\left[\sum Y_i \geqslant 3 \mid H_0\right] = \Pr\left[\mathcal{B}(60; 0.13) \geqslant 3\right] = 0.9883685 > 0.05 = \alpha$$

   por lo que no hay evidencias de que $p > 0.13$. Esto es obvio desde un principio, ya que la estimación muestral $\hat{p} = \dfrac{3}{60} = 0.05$ da la razón (en el sentido de que es más próxima) a H_0 y que "en caso de duda, hipótesis nula"; es este caso, no hay ninguna duda.

   El P–valor podría haberse calculado también mediante

```
binom.test (3, 60, p=0.13, alternative="greater")
```

8. ¿Se puede considerar que la secuencia de datos es aleatoria en el sentido de que no hay tendencia, ni creciente ni decreciente?

   Sean $(x_i)_{i=1}^n$ los datos de la muestra, en el orden en que se han recogido, y $(y_i := i)_{i=1}^n = (1, \ldots, n)$ la secuencia de números naturales que representa la posición de cada dato. Contrastaremos las hipótesis
   H_0: El valor es independiente de la posición $\equiv \rho_{XY} = 0$ frente a
   H_1: Existe tendencia creciente o decreciente según la posición $\equiv \rho_{XY} \neq 0$ donde ρ representa el coeficiente de correlación lineal de Spearman:

```
cor.test (x, 1:n, method="spearman") ## rho = 0.1821415 ; p-value = 0.1637
```

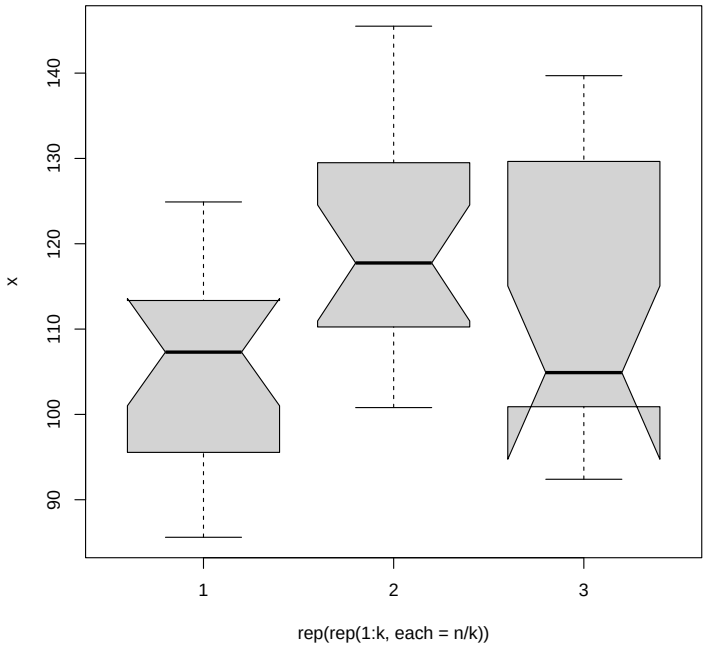
   La correlación muestral es positiva pero baja y, de hecho, no significativamente distinta de cero. Por tanto, se puede considerar que la secuencia de datos es aleatoria en el sentido del enunciado.

   Habría otras posibilidades, como dividir la muestra en dos y comparar si hubo cambio en la tendencia central; o usar alguna versión del contraste χ^2; etc. Por ejemplo, usando Kruskal–Wallis al dividir la secuencia en k grupos:

```
p.valor <- function (k) kruskal.test(x, rep(1:k,each=n/k))$p.value
p.valor(2)  # 0.4118913
p.valor(3)  # 0.0103693
p.valor(5)  # 0.008234161
p.valor(6)  # 0.05119957
p.valor(10) # 0.01537438
p.valor(12) # 0.03967037
```

Hay divisiones que producen rechazo de H_0 , lo que lleva a sospechar que la secuencia no es aleatoria del todo. Hay que tener en cuenta, en todo caso, que se trata de contrastes múltiples y que la representación gráfica no ofrece evidencia clara:

```
k <- 3; boxplot (x ~ rep(1:k,each=n/k), notch=TRUE)
```



Código R

```
x <- c(88.4, 120.4, 99.2, 96.4, 110.1, 110.2, 86.2, 94.7, 108.9, 111.5, 105.6,
       105.3, 106.7, 107.9, 123.4, 121.5, 85.6, 115.2, 124.9, 94, 111.4, 100.8,
       106.4, 104.7, 111.8, 139.6, 130, 145.5, 128, 137.8, 119, 118.7, 108.5,
       122.7, 112.9, 116.8, 144.5, 110.1, 129, 110.4, 104.1, 129.3, 98.7, 118.5,
       111, 103.4, 92.4, 95.2, 105.7, 130, 101.5, 100.6, 133.9, 139.7, 132.2,
       135.5, 113.6, 101.2, 96.4, 104)
t <- range (x)
4/3 * mean(x)
n <- length(x)
alfa <- 0.05
A <- function(alfa1) 1/(1+alfa1^(1/n))-1/(1+(1-alfa-alfa1)^(1/n))
pdf('/tmp/amplitud.pdf')
plot (A, 0, alfa)
dev.off()
cor.test (x, 1:n, method="spearman")
sumYi <- sum (x < 90)
1-pbinom(sumYi-.5, n, .13)
```