3. ARTÍCULOS DE APLICACIÓN

ALGUNAS PARADOJAS Y CURIOSIDADES ESTADÍSTICAS

Carles M. Cuadras
Universidad de Barcelona

1. Introducción

La Estadística observa datos y utiliza el lenguaje matemático para proporcionar modelos, interpretar los resultados y tomar decisiones. Como en otras ramas de la Ciencia, la Estadística no está exenta de contradicciones, paradojas y curiosidades. Puede ser bien o mal utilizada, pero en algunos casos sus resultados confunden o son objeto de curiosidad.

En este trabajo se exponen algunas paradojas y situaciones curiosas, que pueden presentarse en probabilidad, regresión y contraste de hipótesis. Todas ellas tienen cumplida explicación, que se deja a juicio del lector y que se publicará en un próximo boletín.

2. La paradoja de juntar datos

Comenzaremos con un tipo de paradoja bastante maligna, que desconcierta a los usuarios de la Estadística. Se presenta cuando juntamos datos, como en el caso que sigue.

Supongamos un tratamiento contra una enfermedad que consiste en un fármaco aplicado a hombres y a mujeres por separado, para seguidamente juntar los resultados estadísticos obtenidos. En las tres tablas 2×2, exhibidas más abajo, la dependencia entre tratamiento y recuperación es significativa. En los dos primeros casos, los resultados son favorables al fármaco, pues se recuperan el 46 % de los tratados frente al 38 % de los no tratados en el caso de hombres, y el 68 % frente al 58 % en el caso de las mujeres. Pero al juntar las frecuencias, nos encontramos que se recuperan el 49 % de los tratados frente al 54 % de los no tratados. El fármaco es un buen tratamiento para hombres y para mujeres, pero no en general. ¿Por qué?

	Hombres	
	Con tratamiento	Sin tratamiento
Recuperados	700	80
No recuperados	800	130

	Mujeres	
	Con tratamiento	Sin tratamiento
Recuperados	150	400
No recuperados	70	280

	Hombres + Mujeres	
	Con Tratamiento	Sin tratamiento
Recuperados	850	480
No recuperados	870	410

3. Una paradoja del p-valor

Sea $\mathbf{N} = (n_{ij})$ una tabla de contingencia $r \times s$, con frecuencias n_{ij} que combinan r variables filas y s variables columnas. El estadístico V con distribución asintótica ji-cuadrado con $(r-1)(s-1)$ grados de libertad (g.l.), permite decidir si las variables filas son independientes de las variables columnas. El contraste es significativo si $v > \chi^2_\alpha$, donde v es el valor observado y χ^2_α es el valor tabulado para un nivel de significación α y $(r-1)(s-1)$ g.l. Este contraste es equivalente a calcular el $p-$valor $p = P(V > v)$ y decidir que hay dependencia si $p < \alpha$.

Por ejemplo, con $\alpha = 0,05$, en la tabla anterior

del tratamiento a hombres se obtiene $v = 5{,}455$ (1 g.l.), $p = 0,019 < \alpha$, luego es razonable admitir que hay dependencia entre tratamiento y recuperación.

Centrémenos ahora en el grado de significación p. Bajo la hipótesis nula de independencia, la distribución de p es uniforme en $(0, 1)$. Entonces es fácil ver que $-2 \log p$ sigue una ji-cuadrado con 2 g.l. Por lo tanto bastaría comprobar si $-2 \log p$ es significativo. Por ejemplo, $-2 \log(0,019) = 7,926 > 5,991$, siendo $5,991$ el valor tabulado para $\alpha = 0,05$ y 2 g.l. Los dos criterios son equivalentes, pero... ¿es un contraste ji-cuadrado con 1 g.l. o con 2 g.l.?

En general, para una tabla $r \times s$, rechazamos la hipótesis de independencia si:

a) $v > \chi^2_\alpha$, donde $P(V > \chi^2_\alpha) = \alpha$, con $(r-1)(s-1)$ g.l..

b) $-2 \log p > \chi^2_\alpha$, donde $P(\chi^2_2 > \chi^2_\alpha) = \alpha$ con 2 g.l..

En a) y en b) utilizamos el test ji-cuadrado pero trabajamos con distintos grados de libertad. Más en general, todo test ji-cuadrado con m g.l., ¿se reduce a un test ji-cuadrado con 2 g.l.?

Esta aparente paradoja, un poco ingenua, puede confundir al principio, pero tiene una fácil explicación.

Otra paradoja del p-valor

Consideremos los siguientes tamaños muestrales, medias y desviaciones típicas de $p = 2$ variables X (longitud del fémur), Y (longitud del húmero), obtenidas sobre dos poblaciones (Anglo-indios, Indios).

Medias	X	Y
$n_1 = 27$	460,4	335,1
$n_2 = 20$	444,3	323,2
Diferencia	16.1	11.9
Desv. típicas	23,7	18,2

Matriz covarianzas

$$\widehat{\mathbf{S}} = \begin{pmatrix} 561,7 & 374,2 \\ 374,2 & 331,24 \end{pmatrix}$$

Correlación: $r = 0,867$

Estos resultados son debidos a C. R. Rao.

Suponiendo normalidad, los tests t de comparación de medias para cada variable por separado son:

Variable X $t = 2,302$ (45 g.l.) $(p = 0,0259)$,
Variable Y $t = 2,215$ (45 g.l.) $(p = 0,0318)$.

A un nivel de significación $0,05$ se concluye que hay diferencias significativas para cada variable.

Sin embargo, no se obtiene lo mismo aplicando un test con las dos variables conjuntamente. Indicando $\mathbf{d} = (\overline{x}_1 - \overline{x}_2, \overline{y}_1 - \overline{y}_2)'$ obtenemos el estadístico F

$$F = \frac{27 + 20 - 1 - 2}{(27 + 20 - 2)2} \frac{27 \times 20}{27 + 20} \mathbf{d}' \widehat{\mathbf{S}}^{-1} \mathbf{d} = 2,68$$

(2 y 44 g.l.) $(p = 0,079)$,

Por lo tanto ambos tests univariantes son significativos, pero el test bivariante no lo es, contradiciendo este resultado la creencia de que un test multivariante debería proporcionar mayor significación que un test univariante. ¿Qué explicación tiene esta contradicción?

5. Correlaciones que nunca alcanzan el valor uno

Sabemos que el coeficiente de correlación ρ entre dos variable X, Y es un valor que va de -1 a $+1$. Si las variables son normales, la afirmación es correcta, pero si siguen diferente distribución, entonces este rango de valores cambia sensiblemente. Por ejemplo, si X es uniforme e Y es exponencial, cualquiera que sea su distribución conjunta, el coeficiente de correlación verifica

$$-0,866 \leq \rho(X, Y) \leq +0,866.$$

Probar cuál es el valor máximo requiere conceptos relativamente avanzados. Pero es muy fácil argumentar, en este caso, que la correlación entre X e Y no puede alcanzar el valor 1.

Veamos ahora un caso insólito en el que, aparentemente, el coeficiente de correlación puede superar este valor.

6. Una paradoja del coeficiente de correlación

Supongamos que X, Y son dos variables aleatorias definidas sobre la misma población, con covarianza σ_{XY}, variancias finitas σ^2_X, σ^2_Y y coeficiente de correlación de Pearson $\rho = \sigma_{XY}/(\sigma_X \sigma_Y)$.

Sean $X_1, \ldots, X_n$ independientes e igualmente distribuidas como X. Un ejemplo real podría consistir en la estatura Y de un padre y las estaturas $X_1, \ldots, X_n$ de n hijos, donde cada hijo tiene una madre diferente. Nos interesa correlacionar la media $\overline{X}_n$ con Y.

Supongamos $\text{cov}(X_i, Y) = \sigma_{XY}$. La media $\overline{X}_n = (X_1 + \cdots + X_n)/n$ tiene varianza σ_X^2/n y su covarianza con Y es

$$\begin{aligned}\text{cov}(\overline{X}_n, Y) &= \tfrac{1}{n}\text{cov}(\textstyle\sum_{i=1}^n X_i, Y) \\ &= \sigma_{XY}.\end{aligned}$$

El coeficiente de correlación es

$$\begin{aligned}\text{corr}(\overline{X}_n, Y) &= \tfrac{\sigma_{XY}}{(\sigma_X/\sqrt{n})\sigma_Y} \\ &= \sqrt{n}\rho.\end{aligned}$$

Consecuentemente, para n suficientemente grande, el coeficiente de correlación entre la media $\overline{X}_n$ e Y puede ser "mayor" que 1, incluso mucho mayor.

Sorprendentemente, no hay incorrección en el cálculo del coeficiente de correlación $\text{corr}(\overline{X}_n, Y)$. Sin embargo, una correlación no puede superar el valor 1. ¿Qué falla en el planteamiento? ¿Es realmente una paradoja?

7. Una predicción racista

La correlación positiva y alta entre dos variables induce a creer que a un valor alto de una corresponde un valor alto de la otra. Supongamos dos grupos, blancos (B) y negros (N), y que el coeficiente de inteligencia I tenga una correlación alta con una habilidad H (por ejemplo, la capacidad de programar algoritmos). Supongamos que la media de la variable H es la misma en B y en N, pero que la media de la variable I es más alta (por motivos culturales, económicos o sociales) en el grupo B. Una empresa, a efectos de selección de personal, quiere predecir H conociendo I.

Parece razonable afirmar que si un individuo blanco posee el mismo coeficiente de inteligencia I que otro individuo negro, entonces el blanco tendrá un valor esperado H más alto que el negro, puesto que I en el grupo B supera en media al grupo N. Es decir, la predicción H para un mismo I, será más alta en B que en N. Luego el blanco debería ser seleccionado, por ser probablemente más capaz para desarrollar la actividad laboral objeto del empleo.

Sin embargo, este criterio es tendencioso y está equivocado, sucediendo justo al revés, como es fácil comprobar. ¿Por qué?

8. Aumentan las correlaciones simples y disminuye la múltiple

Supongamos que la variable respuesta Y está correlacionada con las variables explicativas $X_1, \ldots, X_k$, que están equicorrelacionadas, es decir,

$$\text{corr}(X_i, X_j) = c \quad i \neq j.$$

Supongamos ahora que las correlaciones simples son

$$\begin{aligned}\text{corr}(Y, X_i) = r \neq 0 \quad &\text{si} \quad i \leq s, \\ \text{cor}(Y, X_i) = 0 \quad &\text{si} \quad i > s.\end{aligned}$$

En otras palabras, Y está correlacionada con las s primeras e incorrelacionada con las otras $k - s$. Indiquemos por $R^2(s)$ el coeficiente de determinación. Se prueba que

$$R^2(s) = \frac{sr^2(1 + (k - s - 1)c)}{(1 + (k - 1)c)(1 - c)}.$$

Es razonable creer que aumentando s también aumente $R^2(s)$. Pero no ocurre así. Por ejemplo, si $r = 0,5$ y $k = 8$, entonces

$$R^2(8) = 0,2844 < R^2(4) = 0,7111.$$

Así pues, resulta que *aumentando* las correlaciones simples entre la variable respuesta y las variables explicativas, *disminuye* el coeficiente de determinación.

Esta curiosidad puede ocurrir en situaciones más generales. Supongamos que Y_1 e Y_2 correlacionan con $X_1, \ldots, X_4$, según $\mathbf{r}_1$ y $\mathbf{r}_2$, respectivamente, siendo $\mathbf{R}$ la matriz de correlaciones entre las x's.

$$\mathbf{r}_1 = \begin{pmatrix} ,6 \\ ,5 \\ ,4 \\ ,3 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & ,3 & ,4 & ,5 \\ ,3 & 1 & ,5 & ,4 \\ ,4 & ,5 & 1 & ,3 \\ ,5 & ,4 & ,3 & 1 \end{pmatrix} \quad \mathbf{r}_2 = \begin{pmatrix} ,6 \\ ,5 \\ ,1 \\ ,1 \end{pmatrix}$$

Se obtiene entonces $R_1^2 = 0,4848 < R_2^2 = 0,7056$. El aumento de dos correlaciones simples ha disminuido el coeficiente de determinación. Esto demuestra que la selección *forward* en regresión múltipe es claramente inapropiada.

Resulta bastante sorprendente que un aumento de las correlaciones simples provoque una disminución de la correlación múltiple. ¿Por qué?

9. Una desigualdad de la correlación múltiple

Si $\mathbf{r} = (r_1, \ldots, r_k)'$ es el vector de correlaciones simples de Y con $X_1, \ldots, X_k$, y las variables explicativas son incorrelacionadas, es decir, $\mathbf{R} = \mathbf{I}$ (matriz identidad), entonces

$$R^2 = r_1^2 + \cdots + r_k^2.$$

En general, se tiene que $\mathbf{R} \neq \mathbf{I}$, siendo razonable suponer que, si las correlaciones simples son positivas, hay redundancia entre las x's, con lo cual

$$R^2 < r_1^2 + \cdots + r_k^2.$$

Sin embargo, la desigualdad contraria

$$R^2 > r_1^2 + \cdots + r_k^2,$$

puede perfectamente ocurrir. Por ejemplo, volviendo a la sección anterior, vemos que

$$R_2^2 = 0,7056 > r_1^2 + \cdots + r_4^2 = 0,63$$

En otras palabras, podemos afirmar que variables correlacionadas no son siempre redundantes, y que mantienen una estructura de dependencia más difícil de interpretar que lo que nos dice la intuición.

10. Mahalanobis mayor que Pearson

Supongamos dos matrices de datos multivariantes $\mathbf{X}, \mathbf{Y}$ resultado de observar k variables, los vectores de medias $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ y la matriz de covarianzas estimada común $\mathbf{S}$.

La distancia de Mahalanobis entre ambas poblaciones es

$$M = (\overline{\mathbf{x}} - \overline{\mathbf{y}})' \mathbf{S}^{-1} (\overline{\mathbf{x}} - \overline{\mathbf{y}}).$$

La distancia de K. Pearson es $K = (\overline{\mathbf{x}} - \overline{\mathbf{y}})' [\mathrm{diag}(\mathbf{S})]^{-1} (\overline{\mathbf{x}} - \overline{\mathbf{y}})$, es decir,

$$K = (\frac{x_1 - y_1}{s_1})^2 + \cdots + (\frac{x_k - y_k}{s_k})^2,$$

siendo $s_1, \ldots, s_k$ las desviaciones típicas de las variables. Podemos entender K como una distancia de Mahalanobis *suponiendo* que las variables están incorrelacionadas.

Ahora bien, la distancia tomando como referencia, ejes ortogonales, debería ser mayor que tomando ejes oblicuos. Existe entonces la creencia de que variables independientes "discriminan" mejor, es decir, que debería verificarse $M < K$. Sin embargo, puede ocurrir lo contrario, es decir, puede ocurrir que Mahalanobis sea mayor que Pearson: $M > K$. ¿Por qué?

Esta curiosidad, también reñida con la intuición, es parecida a la desigualdad anterior con el coeficiente de correlación múltiple R, puesto que $R^2 = \mathbf{r}' \mathbf{R}^{-1} \mathbf{r}$, expresión formalmente similar a la distancia de Mahalanobis.

11. Momentos que no caracterizan la distribución

Los momentos de una variable aleatoria X con función de distribución $F(x)$ y valores en un intervalo (a, b), son las cantidades

$$\alpha_n = E(X^n) = \int_a^b x^n dF(x), \quad n = 0, 1, 2, \ldots$$

Se acepta que la sucesión de momentos $\{\alpha_n\}$ caracteriza la distribución de la variable, en el sentido de que hay una única $F(x)$ con tales momentos. Sin embargo, esto no es en general cierto. Supongamos que $f(x)$ es la densidad de $X = Z^3$, donde Z es normal $N(0, 1/2)$. Consideremos la variable X_ε con densidad

$f_\varepsilon(x) = f(x)\{1 + \varepsilon[\cos\sqrt{3}|x|^{2/3} - \sqrt{3}\sin\sqrt{3}|x|^{2/3}]\}, \quad -1/2 \leq \varepsilon \leq 1/2.$

Entonces se verifica

$$E(X^n) = E(X_\varepsilon^n), \quad n = 0, 1, 2, \ldots$$

a pesar de que $f(x) \neq f_\varepsilon(x)$ para $\varepsilon \neq 0$.

12. Función generatriz que no distingue

Hemos visto que los momentos podrían no determinar la distribución. En cambio, es correcto afirmar que la función generatriz de momentos,

$$M_X(t) = E(e^{tX}) = \int_a^b e^{xt} dF(x),$$

suponiendo que existe, caracteriza totalmente la distribución de X. Sin embargo, ¿es capaz de distinguir bien dos distribuciones distintas la función generatriz de momentos? En teoría sí, pero hay casos en que no se distinguen numéricamente, una de la otra.

Consideremos por ejemplo las funciones de densidad

$$\begin{aligned}\phi(x) &= (2\pi)^{-1/2}e^{-x^2/2},\\ f(x) &= \phi(x)\{1+\tfrac{1}{2}\sin(2\pi x)\}.\end{aligned}$$

La primera es la densidad de X con distribución $N(0,1)$ y la segunda es la de una variable Y que resulta de efectuar una perturbación a X. Ambas funciones son bien distintas, como muestra la Figura 1.

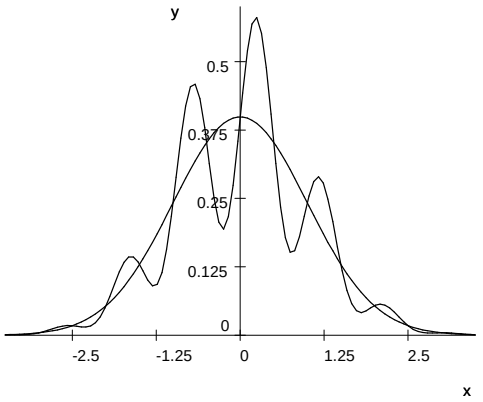

Figura 1: Gráficos de las funciones de densidad $\phi(x)$ y $\phi(x)\{1+\frac{1}{2}\sin 2\pi x\}$, donde $\phi(x)$ es normal estándard.

Las funciones generatrices de momentos son

$$\begin{aligned}M_X(t) &= e^{t^2/2}\\ M_Y(t) &= e^{t^2/2+\log(1+\frac{1}{2}e^{-2\pi^2}\sin(2\pi t))}\end{aligned}$$

que también son distintas. Sin embargo, el gráfico de ambas funciones (Figura 2) las hace prácticamente indistinguibles. No es posible difenciar gráficamente las funciones M_X, M_Y. Por otra parte, en este caso conviene tener en cuenta que los momentos, si no idénticos, son numéricamente muy próximos.

Es fácil construir dos densidades distintas con funciones generatrices muy parecidas. Basta añadir una componente periódica suave. Por ejemplo, la densidad uniforme $f(x) = 1$ y la densidad $g(x) = 1 + \sin(36\pi x)/(36\pi)$, con $0 < x < 1$, son bien distintas pero sus funciones generatrices son casi indistinguibles.

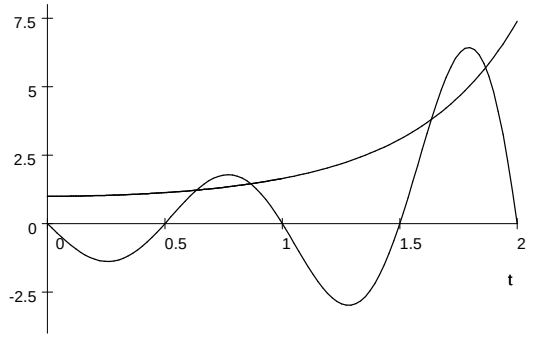

Figura 2: Los gráficos (curva suave de arriba) de $M_X(t)$ y $M_Y(t)$ prácticamente coinciden sobre una misma curva y son indistinguibles numéricamente. Sólo el gráfico (abajo) de $10^9 \times (M_X(t) - M_Y(t))$, con un aumento enorme de la escala, permite apreciar diferencias.

13. La ley de los grandes números falla

Sea X una variable aletoria Poisson con media $\lambda = 1$. Supongamos que $X_1, \ldots, X_n$ son independientes con la misma distribución que X. Sabemos por la ley de los grandes números que $\overline{X}_n$ converge en probabilidad a la media teórica:

$$\overline{X}_n \xrightarrow{P} 1.$$

Es más, en este caso la convergecia de $\overline{X}_n$ a la media $\lambda = 1$ es casi segura, es decir:

$$P(\lim_{n\to\infty} \overline{X}_n = 1) = 1.$$

Sin embargo, un fácil cálculo demuestra que

$$p_n = P(\overline{X}_n = 1) = \frac{e^{-n}n^n}{n!}.$$

Entonces

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} \frac{e^{-n}n^n}{n!} = 0.$$

Por lo tanto, aunque $\overline{X}_n$ converge casi seguramente a 1, $\overline{X}_n$ no puede alcanzar *exactamente* el valor 1 si hacemos tender n a infinito. Obsérvese que incluso es más probable que tengamos $\overline{X}_n = 1$ si n toma un valor pequeño:

n	1	2	5	10	30	50
p_n	$0,367$	$0,270$	$0,175$	$0,125$	$0,072$	$0,056$

La ley de los grandes números aparentementa falla. No obstante, hay una buena explicación para esta anomalía. ¿Cuál?

14. El teorema central del límite falla

El famoso teorema central del límite dice que sumando muchas variables independientes, se obtiene una distribución aproximadamente normal. La suma de unas pocas variables también puede dar normalidad. Por ejemplo, las suma de dos normales independientes es también normal, lo cual es una propiedad de esta ley. La suma de sólo tres variables uniformes ya proporciona una distribución muy parecida a la normal. Nadie discute que la suma de "muchas" variables independientes da lugar a una distribución que sea aproximadamente normal.

Supongamos ahora que $X_1, \ldots, X_n$ son Poisson independientes con parámetro $\lambda = 1/n$. Entonces

$$X = X_1 + \cdots + X_n$$

es una variable aletoria Poisson con media $\lambda = 1$. Tomemos $n = 100$. El valor de n es lo bastante grande como para suponer que se cumpla el teorema central del límite. Sin embargo, la distribución de X está muy alejada de la normal.

¿Falla el teorema central del límite? En realidad no, y por qué no falla se puede explicar fácilmente.

15. Un test de multinormalidad correcto pero poco efectivo

Algunas veces la teoría de la probabilidad confunde al estadístico. Veamos un ejemplo. El teorema de Crámer dice que si $X = X_1 + X_2$ es normal, donde X_1 y X_2 son variables aleatorias independientes, entonces X_1 y X_2 son necesariamente normales. Esta propiedad probabilística, válida para más de dos variables, puede llevar a confusión.

Supongamos ahora que las variables $X_1, \ldots, X_k$ siguen la distribución normal multivariante. Esto ocurre si $X_1, \ldots, X_k$ son combinación lineal de variables $Z_1, \ldots, Z_k$ normales $N(0, 1)$ e independientes.

Consideremos las llamadas componentes principales $Y_1, \ldots, Y_k$, que son combinaciones lineales de $X_1, \ldots, X_k$, incorrelacionadas y por lo tanto independientes. La normalidad de cada una de las componentes principales garantiza la normalidad multivariante de $X_1, \ldots, X_k$, pues estas variables son combinación lineal de $Y_1, \ldots, Y_k$ normales independientes. La normalidad de las componentes principales caracteriza la normalidad conjunta.

Consideremos ahora la suma de las componentes

$$Z = Y_1 + \cdots + Y_k.$$

Según un teorema debido a Cramér, Z es normal si y sólo si $Y_1, \ldots, Y_k$ son variables normales. Luego, dada una muestra multivariante de $X_1, \ldots, X_k$, es decir, una matriz de datos $\mathbf{X}$ de orden $n \times k$, con n grande, bastará obtener las componentes principales y estudiar, mediante un test de bondad de ajuste, la normalidad univariante de Z. Si Z se ajusta a la ley normal, también se ajustan $Y_1, \ldots, Y_k$ y por lo tanto $X_1, \ldots, X_k$ siguen la distribución normal multivariante.

Desde un pùnto de vista probabilístico, el razonamiento parece "correcto". Por ejemplo, considerando los datos de flores del género Iris de R. A. Fisher, especie Setosa, con $n = 50$ y $k = 4$, el test de Kolmogorov de normalidad sobre Z da el resultado $0,079$ con $p > 0,20$ (tabla de Lilliefors), indicando que, sobre la base de una muestra de 50 flores *Iris setosa*, los cuatro caracteres (longitud y anchura de pétalos y sépalos), siguen la distribución multinormal.

Sin embargo, un estadístico sagaz comprobaría, tomando o generando otros datos, que este test de normalidad multivariante falla estrepitosamente. ¿Por qué?

¡En el próximo número se darán las soluciones a las cuestiones planteadas en este artículo!