Algunos problemas de actualidad en estadística*

por

Peter Hall

El desarrollo de nuevas tecnologías ha tenido un gran impacto en los métodos y en el pensamiento estadísticos. No sólo ha cambiado la manera en la que se analizan los datos sino también la naturaleza misma de estos datos. Actualmente se pueden registrar casi de forma continua, mientras que en el pasado luchábamos por conseguir observaciones discretas en instantes muy dispersos. Podemos obtener cantidades ingentes de datos de genotipos sin apenas coste económico y no tenemos que depender únicamente de observaciones de fenotipos tomadas manualmente.

Prácticamente cada tema frontera entre la ciencia y la ingeniería revela nuevos campos de investigación en probabilidad y estadística, y por tanto nuevas oportunidades para los estadísticos. Por ejemplo en finanzas, incluso pequeñas mejoras en el pronóstico de los precios de opciones y valores pueden dar lugar a importantes beneficios. En el ámbito de la minería de datos es frecuente utilizar ideas estadísticas para desarrollar nuevas herramientas como, por ejemplo, los clasificadores, y constantemente se están desarrollando nuevas técnicas.

Hay un considerable potencial de enriquecimiento mutuo entre áreas de aplicación muy diferentes. Por ejemplo, los métodos estadísticos utilizados para predecir el comportamiento intrínsecamente no lineal del movimiento angular de balanceo de una embarcación (figura 1) se han aplicado para predecir otros fenómenos cíclicos, tales como los ciclos financieros. Avances en la tecnología de la información y de las comunicaciones suscitaron nuevos problemas de investigación. En concreto, las técnicas estadísticas basadas en el suavizado, tanto en el dominio del espacio como en el de la frecuencia, tienen su utilidad para extraer información importante a partir de las imágenes hiperespectrales.

Muchas de las técnicas estadísticas actuales se apoyan fuertemente en teoría matemática, bien en su desarrollo o en su justificación. Por ejemplo, existe un rápido y significativo crecimiento en la literatura sobre las aplicaciones de la geometría algebraica para resolver problemas en estadística. En efecto, cuanto más complejos van siendo los problemas para los que se aplican los métodos estadísticos, más apremiante es la necesidad de la intuición que sólo puede proporcionar la estructura geométrica. Así, la geometría algebraica encuentra aplicación en problemas estadísticos paramétricos y no paramétricos, en contextos tan distintos como biología computacional, análisis factorial o inferencia para mixturas de distribuciones.

*Con el título *Some contemporary problems in statistical sciences*, este artículo apareció originalmente en *The Madrid Intelligencer*, Springer (2006), 38–41, publicado con ocasión del *International Congress of Mathematicians* de 2006. *La Gaceta* agradece al autor y a Springer-Verlag la autorización para publicarlo, y a Carmen Nieto Zayas su traducción.

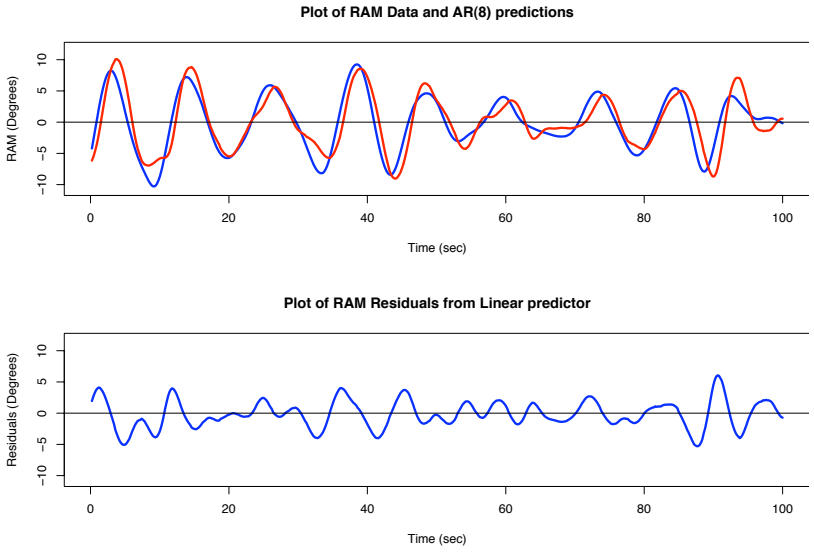

Figura 1: Movimiento angular de balanceo de un barco (RAM: *Roll angular motion*): La curva azul[a] en el cuadro superior es la gráfica del ángulo (en grados) de balanceo de un barco en función del tiempo (en segundos). La curva de color rojo representa la predicción, con tres segundos de anticipación, del ángulo de balanceo mediante un modelo autorregresivo (AR) de orden ocho, utilizando pesos con decrecimento exponencial. Ninguna otra predicción lineal mejora mucho el ajuste. La curva azul en el cuadro de abajo es una gráfica de los residuos de las predicciones, es decir, de la diferencia entre el verdadero valor del movimiento angular y su predicción. A pesar de tratarse de un proceso altamente no lineal presenta bastante estructura.

[a] *Nota de la traductora:* Originalmente, este artículo estaba publicado en color. Si lo ves en blanco y negro, la curva azul es la más oscura, y la roja la más clara.

Una de las formas en las que las nuevas tecnologías están cambiando la naturaleza de la información estadística es al permitir la generación de muestras de bajo coste, donde cada dato corresponde a una curva, a una superficie o a un conjunto de imágenes. Los problemas actuales de clasificación a menudo utilizan datos de este tipo. Por ejemplo, supongamos que tenemos un conjunto de imágenes médicas de cada uno de los individuos para los que se ha diagnosticado que padecen una de k posibles enfermedades, y un conjunto de imágenes de un nuevo paciente cuya situación médica es desconocida y tiene todavía que valorarse. Se quiere determinar cuál, de ser alguna, de las k enfermedades afecta al nuevo paciente, y calcular valores de las verosimilitudes relativas. Como cabe esperar, la reducción de la dimensión juega un papel clave en la resolución de problemas de este tipo, no menor que el de expresar los datos numéricos en función de sus componentes principales. Aquí, como en otros problemas estadísticos actuales (tales como la teoría de la regresión mediante splines penalizadas), la teoría de operadores tiene un papel primordial.

Entre las consecuencias de la gran cantidad de datos que genera la tecnología actual está el creciente número de problemas donde los científicos necesitan valorar muchos análisis diferentes. En los problemas estadísticos clásicos se tiene una única muestra de una población, quizá de tamaño unas decenas de datos, y se desea contrastar una única hipótesis, por ejemplo, la hipótesis de que la media poblacional es cero. Sin embargo, en algunos problemas actuales un investigador dispone de forma rápida y relativamente barata de miles de muestras, cada una conteniendo decenas de datos, y realiza muchos análisis relacionados al mismo tiempo.

En contextos como este, Donoho y Jin [2] discuten distintas técnicas basándose en la idea de Tuckey del «mayor punto crítico», que trata de la significación estadística del número observado de tests estadísticamente significativos. Donoho y Jin estudian explícitamente las diferencias entre problemas de inferencia múltiple, donde hay que estimar muchos parámetros, y otros problemas, mucho menos entendidos, en los que hay que detectar una pequeña proporción de parámetos distintos de cero. Estos últimos problemas se caracterizan por la dispersión y son tratados, por ejemplo, por Johnstone y Silverman [7].

Por supuesto existe una gran variedad de métodos para combinar las conclusiones extraídas en un gran número de análisis. Entre ellos se encuentran los metaanálisis, técnicas de comparaciones múltiples, y métodos FDR (tasa de falsos rechazos o de descubrimientos falsos). Definiremos más adelante lo que se entiende por la tasa de falsos descubrimientos. En particular, el enfoque FDR ha sido motivo de interés en los últimos años, debido en parte a los estudios de Benjamini y Hochberg's [1] para controlar la probabilidad global de error en tests de hipótesis múltiples. Véanse, por ejemplo, [3], [4] y [5].

Reflejando la madurez de la estadística moderna, muchas de las técnicas estadísticas actuales se basan en ideas de hace un siglo. Por ejemplo, la metodología del metaanálisis se puede encontrar en trabajos de Pearson en 1904, y la estadística utilizada en los problemas de análisis múltiples de hoy en día coincide a menudo con la discutida por «Student» en 1908.

Sin embargo, mucho más novedosa es la sofisticación de los argumentos teóricos, extraídos tanto de la teoría de probabilidad como de la estadística teórica, utilizados para sostener nuevas técnicas. Y la dirección y la naturaleza del trabajo científico que motiva los recientes desarrollos podrían venir sólo del siglo XXI. Para subrayar estas características, especialmente la interacción entre la teoría y la aplicación, y para dar solidez a lo que sería, si no, una visión insuficiente, estudiaremos un problema práctico específico que surge frecuentemente.

El problema trata de experimentos con microarrays de genomas, donde uno de los principales objetivos es determinar, para muchos genes diferentes, el nivel de la expresión del gen, es decir, qué gen está expresado de forma diferente («diferenciadamente encendido»). Un solo experimento de microarrays puede producir $N = 10\,000$ muestras, cada una de ellas correspondiente a un gen, y aproximadamente de tamaño $n = 20$.

Cuando se utiliza un microarray de dos canales (véase la figura 2), hasta qué punto el gen i-ésimo (donde $1 \leq i \leq N$) tiene una expresión diferenciada se estudia a través del test unimuestral de la t de Student. Este test se aplica a la i-ésima

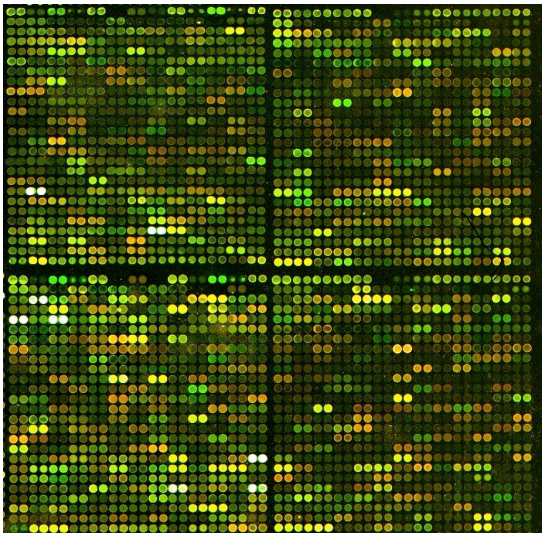

Figura 2: Imagen de colores que representa datos de un microarray de dos canales. Cada punto de la imagen corresponde a un gen, por lo que el i-ésimo punto corresponde al gen i. La imagen representa los datos de un único array, digamos el j-ésimo. El nivel de expresión del gen i en el array j, es decir, el valor de X_{ij}, se deduce del color del i-ésimo punto en el array j, donde los colores varían en la gama de verdes a rojos.

muestra y se utiliza para contrastar la hipótesis nula de que la media de la i-ésima población (o el log ratio) es cero (la hipótesis nula es la hipótesis del «no efecto»). Se dice que se comete un error de tipo I o, equivalentemente, que el resultado del test es un falso positivo, si la hipótesis nula es correcta pero, sin embargo, se rechaza por el test. La proporción de falsos positivos es lo que se llama FDR, la tasa de falsos descubrimientos, o de falsas significaciones. En particular, si N es del orden de decenas de miles, controlar esta tasa requiere muy buenas aproximaciones de la pequeña probabilidad de rechazar la hipótesis nula cuando es cierta.

Para formalizar este problema matemáticamente, sea $\mathcal{X}_i = \{X_{i1}, \ldots, X_{in_i}\}$ la i-ésima muestra, donde se supone que X_{ij} sigue el modelo $X_{ij} = \mu_i + \varepsilon_{ij}$, con $\varepsilon_{i1}, \ldots, \varepsilon_{in_i}$ variables aleatorias independientes e idénticamente distribuidas con media cero. En particular, suponemos que el tamaño de la muestra i-ésima es n_i, dependiendo de i. (En análisis de microarray se tienen tamaños de muestras diferentes, cuando faltan mediciones). Para estos datos se quiere contrastar la hipótesis nula $\mu_i = 0$ que, si es cierta, implica que el i-ésimo gen no está encendido.

Debido a que se utilizará el test N veces, con N grande, y ya que, por hipótesis, queremos un valor de FDR pequeño, se contruye cada test con el objetivo de que se tenga una probabilidad muy pequeña de producir un resultado positivo falso, y una probabilidad moderada de producir al menos un falso positivo si $\mu_i = 0$ para todo i. Esto significa que el test deberá fijar un valor crítico grande, es decir, rechazaríamos

la hipótesis nula sólo si el valor absoluto del estadístico del contraste es bastante grande.

Aunque es razonable suponer, como hicimos anteriormente, que la distribución de X_{ij} es la misma para todo j, con $1 \leq j \leq n_i$, la distribución variará de un valor de i al siguiente, y además será siempre desconocida. Esto hace que sea realmente un reto construir un valor crítico con el que se consiga con seguridad una probabilidad pequeña de producir un resultado positivo falso. Se pretende que el error relativo que se cometa con la aproximación a la probabilidad de rechazo, que es muy pequeña, sea también pequeño, y en esto radica la dificultad.

Por lo general, la suposición de independencia dentro de las filas (esto es, dentro de cada uno de los conjuntos $\mathcal{X}_i$) es razonable, aunque la independencia entre las filas es más cuestionable. Esto no tiene una relación directa con las dificultades discutidas en el párrafo anterior, aunque indica que, en muchas situaciones, se podría tomar una aproximación conservadora, por ejemplo basada en la desigualdad de Bonferroni, para determinar el nivel de significación global de N tests simultáneamente. La desigualdad de Bonferroni simplemente asegura que, para una secuencia de eventos cualesquiera, la probabilidad de la unión no es mayor que la suma de las probabilidades.

El estadístico del test de la t de Student para la i-ésima muestra es de la forma $T_i = (n_i - 1)^{1/2}\, \bar{X}_i / S_i$, donde

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}\,, \qquad S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2\,. \tag{1}$$

La distribución de T_i, bajo la hipótesis nula de que la media poblacional μ_i es cero y, suponiendo que los datos siguen una distribución normal, es una t de Student con $n_i - 1$ grados de libertad. En análisis de microarrays, aunque la distribución de los datos $\mathcal{X}_i$ se aleje de la distribución normal, es frecuente corregirla o «calibrarla», es decir, eligir el llamado valor crítico, usando la t de Student o la distribución normal. La hipótesis nula se rechaza si $|T_i| > x$, donde x es el valor crítico. Por supuesto x depende de n y debería depender también de N si se quiere controlar el número de falsas significaciones.

¿Cómo de graves son los errores que se cometen con las aproximaciones basadas en la distribución normal o en la t de Student? Para plantear esta pregunta de forma más concreta, supongamos que $n_i = n$ para todo i y que se quiere que la probabilidad de rechazar una sola hipótesis nula de entre las N converja, si cada media μ_i es cero, a un valor fijo $\alpha \in (0, 1)$ cuando N y n aumentan. Es frecuente tomar $\alpha = 0{,}05$. Tomando α fijo y estrictamente menor que 1, aseguramos que la probabilidad de rechazar una hipótesis nula, cuando las N hipótesis son ciertas, es en el límite menor que 1. Con este objetivo, ¿cómo de grande debe ser N, como función de n, si calibramos bajo la hipótesis ficticia de que T_i sigue una distribución t de Student o una distribución normal?

La respuesta es que se puede tomar N tan grande como $N = \exp\{o(n^{1/3})\}$ (es decir, $\log N = o(n^{1/3})$), pero no mayor, salvo que se pueda asumir que la distribución de los datos tiene una simetría considerable. Por otro lado, si aproximamos

utilizando métodos empíricos que estiman con precisión los efectos de las distintas distribuciones de los datos en diferentes filas, en vez de emplear las convencionales t de Student o aproximación normal, se puede ir sustancialmente más lejos que $N = \exp\{o(n^{1/3})\}$ sin preocuparse por la simetría. Estos resultados son ciertos si no suponemos nada sobre la independencia de las muestras $\mathcal{X}_i$ y calibramos utilizando las cotas de Bonferroni; además siguen siendo ciertos si se cumple la hipótesis de independencia y lo aprovechamos.

El hecho de que N pueda tomarse exponencialmente grande, como función del tamaño de muestra n, justifica la aproximación mediante la t de Student o la distribución normal, siempre que N no sea excesivamente grande. Sin embargo, son pocas las situaciones con estas características estudiadas por los investigadores; por ejemplo $N = 10\,000$ y $n = 20$, deberían ser una excepción. En particular, si $N = 10\,000$ y $n = 20$ entonces $(\log N)/n^{1/3}$ no es pequeño, sino aproximadamente igual a 3,4. Por tanto la aproximación a la t de Student o la normal no es aconsejable. En tales casos deberíamos ser menos ambiciosos, utilizando un valor crítico x menor y aceptando en consecuencia una mayor tasa de falsos positivos (FDR).

Sin embargo, hay otras alternativas para este análisis. Son útiles en los frecuentes casos donde n es bastante pequeño e incluye un «suavizado estadístico». En nuestro caso, esto concierne a información compartida entre los genes.

Los resultados presentados anteriormente se adaptan a situaciones con grandes dispersiones utilizando estadísticos como la media estudentizada («estudentizar» hace referencia a la operación de estandarizar la escala, utilizando la desviación típica muestral, que en nuestro caso es S_i en (1)). Los desarrollos con desviaciones grandes están relacionados con errores cometidos en aproximaciones de las probabilidades en las colas de las distribuciones, tales como aquellos que se alcanzan mediante el teorema central del límite:

$$\frac{\Pr_0(T_i > x)}{\Pr(Z > x)} \to 1 \qquad \text{o} \qquad \frac{\Pr_0(T_i > x)}{\Pr(Z > x)} \exp\left(c_i\, x^3\, n^{-1/2}\right) \to 1 \tag{2}$$

cuando $x = x(n) \to \infty$ y $n \to \infty$ a la vez, donde $\Pr_0$ indica la probabilidad bajo la hipótesies nula $\mu_i = 0$, Z denota una variable aleatoria con distribución normal estándar, y c_i es una constante relacionada con la asimetría de la distribución de X_{i1} (se sigue suponiendo que $n_i = n$).

Shao [8] y Jing et al. [6] prueban que, siempre que x aumente más lentamente que $n^{1/6}$ y bajo la hipótesis de que los valores absolutos de los momentos de orden tres estandarizados de X_{i1} están uniformemente acotados, la primera parte de (2) converge uniformemente en i (los momentos estandarizados son los momentos de X_{i1} después de corregir la localización y la escala). Wang [9] demuestra que, mientras x sea de orden menor que $n^{1/4}$, la segunda parte de (2) se cumple uniformemente en i, bajo la hipótesis de que los valores absolutos de los momentos de orden cuatro estandarizados estén acotados uniformemente.

Los resultados de Shao [8], Jing et al. [6] y Wang [9] son llamativos por su generalidad, ya que no se satisfacen fórmulas del tipo (2) para distribuciones de sumas de variables aleatorias independientes bajo condiciones tan débiles como las supuestas sobre los momentos. Los resultados, así como algunas propiedades relacionadas,

sugieren que la media estudentizada es mucho más robusta frente a distribuciones muestrales de colas pesadas que la media no estudentizada.

Por lo tanto, el desarrollo de métodos estadísticos para resolver una gran cantidad de problemas prácticos, incluyendo el control de la tasa de falsos positivos, son una ayuda significativa en la teoría de la probabilidad y en la estadística. Estas conexiones demuestran que las nuevas tecnologías, en materias tan diversas como la biología, ingeniería y las finanzas, proponen nuevos retos a muchas áreas de la Estadística.

Agradecimientos. Agradezco a Jianqing Fan, Iain Johnstone, Gordon Smythe, Qiying Wang y Sue Wilson sus útiles comentarios. Kirk Hampel y Gordon Smyth realizaron las figuras 1 y 2, respectivamente.

Referencias

[1] Y. Benjamini y Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B* **57** (1995), 289–300.

[2] D. L. Donoho y J. Jin, Higher criticism for detecting sparse heterogeneous mixtures, *Ann. Statist.* **32** (2004), 962–994.

[3] B. Efron, Large-scale simultaneous hypothesis testing: the choice of a null hypothesis, *J. Amer. Statist. Assoc.* **99** (2004), 96–104.

[4] J. Fan, Y. Chen, H. M. Chan, P. Tam y Y. Ren, Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells, *Proc. Nat. Acad. Sci. USA* **102** (2005), 17751–17756.

[5] J. Fan, H. Peng y T. Huang, Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. (With discussion.), *J. Amer. Statist. Assoc.* **100** (2005), 781–813.

[6] B. Y. Jing, Q.-M. Shao y Q. Y. Wang, Self-normalized Cramér type large deviations for independent random variables, *Ann. Probab.* **31** (2003), 2167–2215.

[7] I. M. Johnstone y B. W. Silverman, Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences, *Ann. Statist.* (2004) **32**, 1594–1649.

[8] Q.-M. Shao, A Cramér type large deviation result for Student's t statistic, *J. Theoret. Probab.* **12** (1999), 385–398.

[9] Q. Wang, Limit theorems for self-nomalized large deviation, *Electronic J. Probab.* **38** (2005), 1260–1285.

Peter Hall, Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, 3010, Australia
Correo electrónico: halpstat@ms.unimelb.edu.au

Traducido por Carmen Nieto Zaya, E. U. de Estadística, Universidad Complutense de Madrid, Avda. Pta. Hierro s/n, 28040 Madrid