2. ARTÍCULOS DE APLICACIÓN

QATRIS IMANAGER, UN SISTEMA CBIR BASADO EN REGRESIÓN LOGÍSTICA

José P. Arias Nicolás[1], **Fernando Calle Alonso**[2] **e Inés M. Horrillo Sierra**[2]
[1] Departamento de Matemáticas
Universidad de Extremadura
[2] Departamento de I+D+i
SICUBO

1. Introducción

En los últimos años el interés por el mundo de las imágenes digitales ha crecido enormemente. Esto ha llevado a la aparición de nuevos campos de investigación, como es el de la Recuperación de Imágenes Basado en Contenido (CBIR). El objetivo de las técnicas CBIR es la recuperación de información basada en las características extraídas de las imágenes que se encuentran en la base de datos, para encontrar, de la forma más eficiente, aquellas que son similares a una dada. En este trabajo desarrollamos un método de comparación por pares basado en regresión logística bayesiana para determinar imágenes similares cuando se dispone de cierta información sobre la similitud que existe entre algunas de las imágenes de la colección. El procedimiento está basado en un método de agregación de preferencias dado por Arias-Nicolás et al. [2] para determinar una función de utilidad aditiva para un grupo de decisores a partir de sus funciones de utilidad individuales. Este método es particularmente útil en problemas con un gran número de características y pocas imágenes.

Las herramientas de búsqueda de información en la red como Google o Yahoo representan hoy en día los principales soportes de búsqueda de información. Sin embargo, estos motores de búsqueda tienen capacidades muy limitadas en cuanto a la determinación de la semántica de los documentos recuperados y en la adecuación a las necesidades concretas de información del usuario.

En los clásicos sistemas CBIR, los resultados de la búsqueda son un conjunto de imágenes con características similares a la imagen de referencia. Sin embargo imágenes con un alto índice de características similares a la imagen de referencia pueden ser muy diferentes desde el punto de vista del usuario que realiza la búsqueda. Esta discrepancia es conocida como "semantic gap". En este trabajo, introducimos un novedoso método para la recuperación de imágenes basadas en contenido, que utiliza la información que proporcionan las características y la opinión del usuario que realiza la búsqueda. El método ha sido implementado en "Qatris IManager", un potente buscador, editor y clasificador de imágenes perteneciente a la empresa SICUBO, S.L.; empresa de base tecnológica, nacida en el seno de la Universidad de Extremadura y ubicada en Cáceres. "Qatris IManager", puede realizar consultas atendiendo a tres tipos de características: color, textura y forma, Chorás [4]. En este trabajo describimos el algoritmo de búsqueda considerando únicamente características de color.

2. Metodología

El objetivo del trabajo es determinar un buscador de imágenes similares, de acuerdo a las características de color extraídas, que pueda ayudar al decisor a clasificarlas. Todo buscador consta de dos fases, una de entrenamiento y otra posterior de prueba. En la primera fase, el sistema requiere un conjunto de imágenes que empleará para determinar un patrón de búsqueda inicial que permita clasificar nuevas imágenes en una u otra clase. El entrenamiento, y por tanto el buscador resultante, será supervisado, es decir, necesita de un conjunto de imágenes de los que se conoce su clase a priori. Este tipo de entrenamiento precisa por tanto, una fase de preclasificación efectuada normalmente por un experto que asigne cada imagen a una de las clases. De esta forma, el sistema conoce cuáles son las clases disponibles y además, dispone de un conjunto de ejemplos de cada una de ellas.

2.1. Obtención del vector de características de color

La extracción de las características de color se lleva a cabo con el modelo de color HSV (Hue-Saturation-Value). Para ello se ha hecho necesario la utilización de un algoritmo de conversión del modelo RGB (Red-Green-Blue) a HSV. Ambos, HSV y RGB son los espacios de color más utilizados, ya que son modelos que representan el color en términos de valores de intensidad. El modelo de color RGB está compuesto por los colores primarios Rojo, Verde y Azul. Este sistema define el modelo que se utiliza por ejemplo en los monitores CRT o en las pantallas TFT. La suma de distintas cantidades de estos colores primarios da lugar al color deseado. El modelo de color HSV está basado en los utilizados en pintura (matiz, sombra y tono), Figura 1. El sistema de coordenadas es un cono hexagonal donde los valores representan la intensidad de un color. El matiz y la saturación están íntimamente relacionados con la manera que tenemos los humanos de percibir los colores.

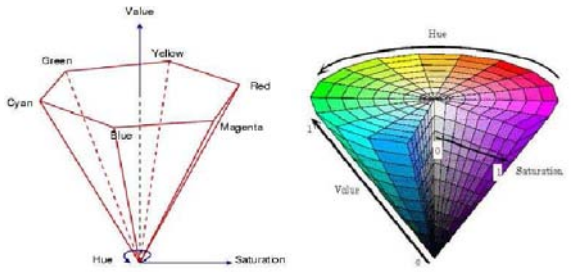

Figura 1. Modelo HSV

La conversión del modelo RGB a HSV se ha llevado a cabo debido a que el número de colores para la extracción de características se ha reducido pasando de los 2^{24} colores iniciales a 14. Para llevar a cabo esta discretización el modelo HSV es mucho más intuitivo. En un primer paso, la imagen se carga en una estructura RGB, es decir una matriz de dimensiones alto, ancho y RGB, donde cada valor oscila entre 0 y 255. Esta matriz RGB se transforma en otra HSV y, troceando el cono discreto, se obtienen los 14 colores finales.

2.2. Método de comparación por pares

Una vez que hemos obtenido el vector de características de cada imagen de la base de datos, aplicamos el método propuesto. Supongamos que cada imagen está representada por un grupo de n características y denotemos por $\mathcal{A}$ al conjunto de imágenes y por d_i a la función distancia que modela la discrepancia o disimilitud entre imágenes con respecto a la característica i-ésima, $i = 1, ..., n$.

El objetivo es determinar una distancia d_g que represente la discrepancia entre imágenes con respecto a todas las características conjuntamente. Teniendo una clasificación a priori de imágenes o las opiniones a priori de un experto acerca de la similitud entre las mismas, nuestro objetivo es encontrar las imágenes similares a una dada y con ello ayudar al decisor a clasificar nuevas imágenes. En primer lugar, se muestrea un conjunto de r pares de imágenes de la base de datos. Para cada par de imágenes $j = (a, b)$, $j = 1, \ldots, r$, calculamos:

1. Las variables independientes, i.e., la diferencia entre sus características:

$$\mathbf{x}_j = (d_1(a,b), d_2(a,b), ..., d_n(a,b)).$$

2. La variable respuesta o dependiente, Y_j (Y_{ab}), que toma el valor 0 si las imágenes son similares para el usuario, ó 1, si no lo son.

Si no disponemos de la opinión del experto, pero tenemos una clasificación previa de las imágenes en carpetas, la variable respuesta toma el valor 0 si las imágenes pertenecen a la misma clase, ó 1, si no. Si las imágenes que se encuentran en las mismas carpetas no son similares para el usuario, Qatris IManager realiza una clasificación automática, de acuerdo a la distribución espacial de las características extraídas de las imágenes. Esta clasificación se realiza previamente al muestreo de pares. Esta es la fase de entrenamiento del sistema.

A continuación, se aplica regresión logística al vector de datos $(x_{j1}, x_{j2}, \ldots, x_{jn}, y_j)$, $j = 1, \ldots, r$. Regresión logística forma parte de los llamados Modelos Lineales Generalizados (GLM). En esta clase de modelos se incluyen tanto modelos clásicos de regresión y ANOVA, como otras interesantes y útiles técnicas multivariantes. El lector interesado en GLM puede consultar, entre otros, los textos de Agresti [1], McCullagh y Nelder [6] y Ryan [8].

Ya que nuestro objetivo es determinar una medida de discrepancia entre imágenes y las variables independientes son no negativas, el predictor lineal Y será no negativo. Por lo tanto consideraremos

una función de enlace que transforme la probabilidad de que las imágenes sean diferentes, π en un valor del intervalo $[0, +\infty)$. Entonces, consideramos la función de enlace:

$$g(\pi) = \log\left(\frac{1+\pi}{1-\pi}\right).$$

Uno de los principales problemas de la regresión logística clásica es el sobreajuste (overfitting) de los datos. Un sobreajuste significa tener más variables explicativas de las que realmente necesita la variable dependiente en el conjunto de entrenamiento. Como consecuencia, al aplicar un modelo de regresión clásico, los parámetros estarán afectados por ese sobreajuste y su significado es incorrecto, por lo que se dice que el modelo está sobreajustado.

El enfoque bayesiano soluciona este problema asignando una distribución a priori para que los parámetros se mantengan en un intervalo limitado, normalmente cercano a cero. Además, a los parámetros asociados a las variables "sobrantes" es posible asignarles un valor nulo para que no afecten al modelo y se corrija el sobreajuste de los datos iniciales.

La regresión logística binomial es un modelo de probabilidad condicional de la forma:

$$P\left[Y = 1|\mathbf{x}, \mathbf{B}\right] = \frac{e^{\mathbf{x}^t\mathbf{B}} - 1}{e^{\mathbf{x}^t\mathbf{B}} + 1},$$

parametrizado por el vector $\mathbf{B} = (\beta_1, ..., \beta_n)$. Cada elemento del vector $\mathbf{B}$ es un parámetro que está asociado a una característica y representa la importancia que tiene ésta en la búsqueda de imágenes similares. Para determinar las imágenes similares a una nueva, nos basamos en el vector de probabilidades condicionales estimadas por el modelo. Para ello, simplemente se ordenan las imágenes de menor a mayor probabilidad estimada.

Consideramos la siguiente muestra de pares de imágenes con su correspondiente variable respuesta, $\mathbf{D} = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_r}, y_r)\}$. La estimación máximo verosímil de los parámetros $\mathbf{B}$ es equivalente a minimizar el opuesto del modelo logit:

$$l(\mathbf{B}|\mathbf{D}) = -\sum_{i=1}^{r} y_i \log\left(0{,}5(e^{\mathbf{x}_i^t\mathbf{B}} - 1)\right) -$$

$$\sum_{i=1}^{r} \log\left(0{,}5(e^{\mathbf{x}_i^t\mathbf{B}} + 1)\right)$$

Como comentábamos antes, es importante corregir el sobreajuste del conjunto de entrenamiento en un modelo de regresión logística para permitir predicciones más precisas de nuevos elementos. Una alternativa es utilizar una distribución a priori de $\mathbf{B}$ que proporcione una alta probabilidad de que la mayoría de los elementos de $\mathbf{B}$ sean 0 o estén cercanos a 0. La aproximación bayesiana del modelo de regresión logística más ampliamente utilizada es la de imponer una distribución gaussiana para cada parámetro β_j.

Una vez estimado el valor de $\mathbf{B}$, obtenemos, para cada par de imágenes a y b:

$$\hat{\pi}_{ab} = \hat{P}[Y_{ab} = 1] = \frac{e^{\mathbf{x}_{ab}^t\hat{\mathbf{B}}} - 1}{e^{\mathbf{x}_{ab}^t\hat{\mathbf{B}}} + 1},$$

la probabilidad de que a y b sean diferentes.

Obsérvese que para todo imagen $a \in \mathcal{A}$, se cumple que $d_i(a, a) = 0,\ i = 1, \ldots, n$. Por lo tanto $\hat{\pi}_{aa} = 0$. Así que, dadas dos imágenes $a, b \in \mathcal{A}$, decimos que la imagen a es similar a la imagen b para el usuario, si $\hat{\pi}_{ab}$ es cercano a 0. Cuando $\hat{\pi}_{ab} = 0$, diremos que las imágenes son iguales. Sin embargo, si dos imágenes a y b son muy diferentes con respecto a alguna característica, $d_i(a, b) \to +\infty$ y $\beta_i > 0$, entonces la probabilidad de que a y b sean diferentes, $\hat{\pi}_{ab} \to 1$.

Así que, para una nueva imagen $\mathbf{c}$, podemos computar $\hat{\pi}_{ca_j}$, que puede ser interpretado como el grado de discrepancia entre la imagen c y la imagen a_j. Nótese que si c y a_j son muy similares, la probabilidad de discrepancia es cercana a 0. Entonces, debemos buscar las imágenes a_j tales que la probabilidad de que ésta y c sean diferentes sea cercana a 0.

- Teniendo en cuenta que $\hat{\pi}_{ab}$ cercano a 0, es equivalente a $\mathbf{x}_{ab}^t\hat{\beta}$ cercano a 0, podemos considerar
$$d_g(\cdot) = \sum_{i=1}^{n} \hat{\beta}_i d_i(\cdot),$$
como una medida de discrepancia entre imágenes.

- Si no podemos comparar todas las imágenes, pero podemos hacer comparaciones por pares para algunas imágenes, el método que proponemos es también aplicable, así pues, no es necesario conocer la opinión del usuario para todos los pares de la base de datos. Cuanta más información tengamos, más precisos serán los resultados. De acuerdo a la fórmula de

Freeman ([5]), debemos disponer de, al menos, $10(n+1)$ observaciones, siendo n el número de características. Perduzzi et al. [7] consideran que es suficiente disponer de 10 observaciones o pares de imágenes por variable.

- No es necesario conocer las distancias individuales. El método tambien es aplicable, definiendo:

$$x_{ab}^{i} = \begin{cases} 0, & \text{si } a \text{ y } b \text{ son similares} \\ & \text{con respecto a } i, \\ 1, & \text{en caso contrario,} \end{cases}$$

para $i = 1, 2, \ldots, n$. Es decir, sólo necesitamos conocer si dos imágenes son similares o no, respecto a cada una de las características.

- La relación de orden, $\preceq_a$ inducida por el método de regresión logística para cada una las imágenes $a \in \mathcal{A}$;

$$a_k \preceq_a a_l \iff d_g(a, a_l) \geq d_g(a, a_k) \qquad (1)$$

es comparable, reflexiva, antisimétrica y transitiva. Es decir, $\preceq_a$ define un orden débil de preferencia en $\mathcal{A}$.

2.3. Aprendizaje

La Figura 2 muestra el resultado de una búsqueda en QATRIS IManager.

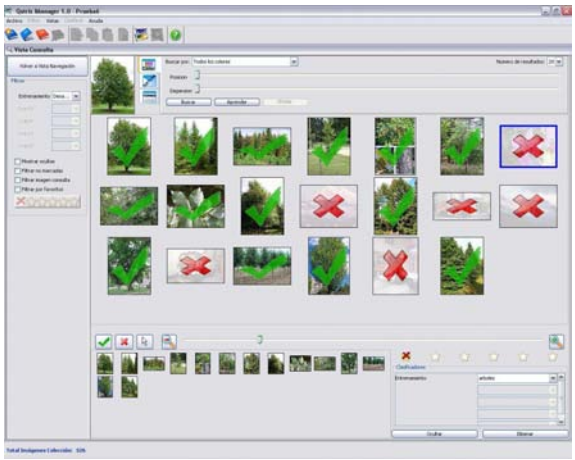

Figura 2. Qatris IManager

El sistema muestra las imágenes de la base de datos más parecidas a la imagen de consulta (IC). El usuario puede interaccionar con el sistema señalando, cuáles de las imágenes mostradas se corresponden con lo que el usuario está buscando (IR-OK) y cuáles no (IR-NO OK). Con esta información, mediante un proceso de feedback o retroalimentación, se permite incorporar la opinión del usuario en cada consulta, añadiendo nuevos pares, Tabla 1.

Imágenes	IC	IR-OK	IR-NO OK
IC	...	$Y = 0$	$Y = 1$
IR-OK	$Y = 0$	$Y = 0$	$Y = 1$
IR-NO OK	$Y = 1$	$Y = 1$	...

Tabla 1. Aprendizaje en Qatris IManager.

Si en la consulta el usuario marca una imagen como IR-OK y otra como IR-NO OK, el sistema considera estas imágenes distintas ($Y = 1$). Si dos imágenes son marcadas como IR-OK se consideran similares entre sí ($Y = 0$). Las imágenes mostradas por el sistema, que no se corresponden con lo que el usuario busca (IR-NO OK) no constituyen un nuevo par de información, es decir no se añadirían al conjunto $\mathbf{D}$. Tampoco proporcionan información las imágenes no marcadas por el usuario.

El sistema aprende de las respuestas proporcionadas por el decisor y actualiza el vector $\mathbf{B}$ con dicha información. Utilizando metodología bayesiana, se obtiene la distribución a posteriori de los parámetros, [3]. Una vez optimizada ésta, el sistema ofrece al decisor nuevas imágenes similares. El proceso se repite hasta que el decisor se muestra conforme con el resultado.

3. Interfaz de Qatris IManager

Consideramos una colección de imágenes (cedidas por SICUBO, S.L.) clasificadas en 7 grupos de imágenes similares: árboles, atardeceres, imágenes en blanco y negro, bomberos, flores, nieve, objetos y otros. La Figura 3 muestra una secuencia de las etapas que constituyen el proceso de búsqueda y aprendizaje en Qatris IManager.

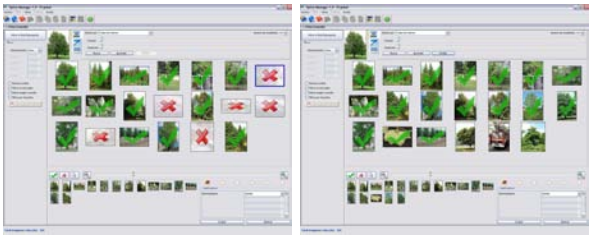

Figura 3. Búsqueda en Qatris IManager

En la primera de ellas se procede al cargado de imágenes en el sistema. En esta fase, el sistema extrae automáticamente el vector de características (color, textura y forma) de cada una de las imágenes de la colección. Tras un muestreo de pares de imágenes, se obtiene una estimación inicial de la distribución de los parámetros, que da lugar a una primera medida de discrepancia entre imágenes. El sistema muestra, en este ejemplo, las 20 imágenes más parecidas a la imagen de referencia (en este caso, la imagen de un árbol). Las siguientes secuencias muestran las etapas de interacción con el usuario y obtención de nuevos resultados. Una vez que el usuario selecciona las imágenes que se ajustan a lo que está buscando y las que no, el usuario tiene la posibilidad de obtener nuevos resultados, pulsando:

- BUSCAR: No se actualizan los pesos.
- APRENDER: Se actualizan los pesos con la informacion proporcionada por el usuario.
- OLVIDAR-BUSCAR: El sistema olvida toda la información proporcionada por el usuario en las consultas anteriores, inicializando el conjunto de pares **D**.

En todo caso, el sistema no volverá a mostrar las imágenes marcadas como IR- NO OK por el usuario. Tras la búsqueda y, atendiendo a los resultados obtenidos, el usuario puede clasificar la imagen de consulta en una de las clases existentes o bien dejar al sistema que realice la clasificación de forma automática.

Agradecimientos

Nuestro más sincero agradecimiento al departamento de I+D+i de SICUBO por su inestimable ayuda y por proporcionarnos las imágenes que han hecho posible la realizacion de esta trabajo. Los autores tambien quieren agradecer el apoyo del Centro para el Desarrollo Tecnológico Industrial y del Ministerio de Educación y Ciencia a través de los proyectos NEOTEC-20050367 y TSI2004-06801-C04-03, respectivamente.

Referencias

[1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley.

[2] Arias-Nicolás, J.P., Martín, J. and Pérez C. (2007). A logistic regression-based pairwise comparison method to aggregate preferences. *Group Decision and Negotiation*, **por aparecer**.

[3] Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer.

[4] Chorás, R.S. (2003). Content-Based Retrieval Using Color, Texture and Shape Information. *Lecture Notes in Computer Sciences*, **2905**, 619-626.

[5] Freeman, D.H. (1987). *Applied categorical data analysis*, New York Marcel Dekker Inc.

[6] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall.

[7] Perduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373-1379.

[8] Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.

RESPUESTAS A ALGUNAS PARADOJAS Y CURIOSIDADES ESTADÍSTICAS

Carles M. Cuadras
Universidad de Barcelona

1. Introducción

En este trabajo se exponen las soluciones a algunas paradojas y situaciones curiosas, que pueden presentarse en probabilidad y estadística, publicadas en el *Boletín de la SEIO*, 23 (1), 24-29, véase Cuadras (2007). Las soluciones que aquí se proponen no son necesariamente las únicas posibles.

2. La paradoja de juntar datos

En la primera paradoja nos encontrábamos con la sorpresa de que un tratamiento eficaz para hombres y mujeres por separado, en el sentido de que mejoran más los tratados que los no tratados, los resultados se invierten si juntamos las dos tablas 2×2. Concretamente, 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 de hombres y mujeres, resultan que se recuperan el 49 % de los tratados frente al 54 % de los no tratados. Las tablas de datos y la paradoja aparecen en Székely (1986, p. 135), aunque el autor no proporciona ninguna solución.

Una explicación para esta paradoja, conocida como paradoja de Simpson, es como sigue. Al considerar la tabla con todas las frecuencias, estamos mezclando dos poblaciones distintas, con proporciones significativamente diferentes en cuanto a recuperación en hombres y mujeres. Hablando en términos de probabilidades, donde "Rec" significa paciente recuperado y "Trat" que ha recibido tratamiento, podemos escribir

$$P_H(\text{Rec/Trat}) = 0,46 \qquad P_M(\text{Rec/Trat}) = 0,68$$

Hay en total 2610 personas y las proporciones de hombres y mujeres son $P(\text{H}) = 0,65$, $P(\text{M}) = 0,35$. Entonces la probabilidad de recuperarse si ha seguido tratamiento, haciendo abstracción de que sea hombre o mujer, es

$$P(\text{Rec/Trat}) = P_H(\text{Rec/Trat})P(H) + P_M(\text{Rec/Trat})P(M) = 0,54$$

Por otra parte, la probabilidad de recuperarse si no ha seguido tratamiento es $P(\text{Rec/SinTrat}) = 0,46$. Es decir, el 54 % se recuperan frente al 46 % que no se recuperan, y la superioridad del tratamiento se confirma al juntar los datos de hombres y mujeres. Evidentemente, si una persona que ha seguido el tratamiento se interesa por la probabilidad de recuperarse, sabremos de entrada si es un hombre o una mujer. Pero si por alguna razón esta información no está disponible, la probabilidad debe calcularse ponderando con las proporciones de hombres y mujeres. En cuanto a la significación estadística de la influencia del tratamiento en la mejora de los pacientes, puesto que tenemos dos tablas 2×2 independientes, debería aplicarse el test de Mantel-Haenszel. Véase Lee (1992).

3. Solución a la primera paradoja del p-valor

Si V es un estadístico de contraste en un test ji-cuadrado con m grados de libertad, bajo la hipótesis nula, el $p-$valor $p = P(V > v)$ sigue la distribución uniforme en el intervalo $(0, 1)$. Se toma la misma decisión tanto si $v > \chi^2_\alpha$ como si $p < \alpha$, donde α es el nivel de significación. Pero $-2 \log p$ sigue una ji-cuadrado con 2 g.l., y por lo tanto podemos plantear el test utilizando $-2 \log p$. Para $m \neq 2$ resulta paradójico o incoherente que un contraste ji-cuadrado con m g.l. se convierta en uno con 2 g.l.

En realidad, cualquier variable continua X con función de distribución F se puede convertir en una ji-cuadrado con 2 g.l. Basta tomar $-2 \log(F(X))$. En particular, cualquier estadístico (test F por ejemplo), se puede reducir a un ji-cuadrado con 2 g.l. siguiendo el mismo procedimiento. No se trata pues de una paradoja, sino de un simple cambio de variable que aparentemente modifica el estadístico o los grados de libertad.

4. Solución a la segunda paradoja del p-valor

Esta paradoja aparece en Rao (1952, p. 252). Se obtenían dos t de Student univariantes significativas para dos variables x, y, por separado y una F

de un test bivariante no significativa:

$$\begin{array}{ll} x & t = 2,302 \quad (45 \text{ g.l.}) \quad (p = 0,0259), \\ y & t = 2,215 \quad (45 \text{ g.l.}) \quad (p = 0,0318). \\ (x,y) & F(2,44) = 2,68 \quad (p = 0,078) \end{array}$$

¿Cómo se explica que los dos tests univariantes sean significativos pero el bivariante no? Vamos a dar una explicación que seguramente no es la única posible.

Interpretemos geométricamente esta paradoja. Con nivel de significación 0,05, y aplicando el test T^2 de Hotelling, aceptaremos la hipótesis nula bivariante si el vector diferencia $\mathbf{d} = (x\ y)'$ pertenece a la elipse

$$\frac{n_1 n_2}{n_1 + n_2} \mathbf{d}' \begin{pmatrix} 561,7 & 374,2 \\ 374,2 & 331,24 \end{pmatrix}^{-1} \mathbf{d} \leq 3,2,$$

donde 3,2 es el punto crítico para una F con 2 y 44 grados de libertad. Así pues no hay significación si x, y verifican la inecuación

$$0,04\,036\,9x^2 - 0,0912\,1xy + 0,06845\,6y^2 \leq 3,2.$$

Análogamente, en el test univariante y para la primera variable x, la diferència $d = \overline{x}_1 - \overline{x}_2$ debe verificar

$$|\sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\frac{d}{s_1})| \leq 2,$$

siendo 2 el valor crítico para una t con 45 g. l. Procederíamos de forma similar para la segunda variable y. Obtenemos así las cuatro rectas

Variable x: $0,143x = \pm 2$, Variable y: $0,1862y = \pm 2$.

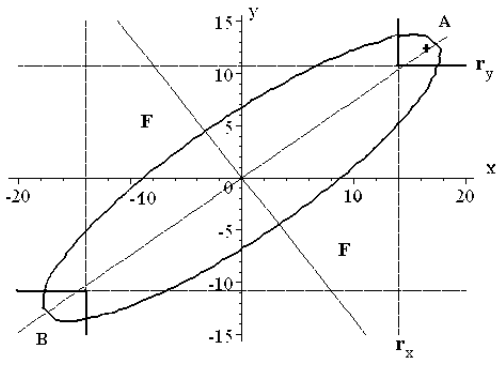

Figura 1. Un test de comparación de poblaciones bivariante puede ser menos significativo que dos tests univariantes.

En la figura 1 podemos visualizar la paradoja. Los valores de la diferencia que están a la derecha de la recta vertical $\mathbf{r}_x$ son significativos para la variable x. Análogamente los que están por encima de la recta horizontal $\mathbf{r}_y$ lo son para la y. Por otra parte, todos los valores que están fuera de la elipse (región **F**) son significativos para las dos variables. Hay casos en que x, y por separado no son significativos, pero conjuntamente sí. No obstante, existe una pequeña región por encima de $\mathbf{r}_y$ y a la derecha de $\mathbf{r}_x$ que cae dentro de la elipse. Para los datos del ejemplo, se obtiene el punto señalado con el signo $+$, para el cual x e y son significativas pero no (x, y). Así x e y son significativas si el punto se encuentra en el cuadrante **A**. (Una simetría con respecto al origen nos permitiría considerar otras dos rectas y la región **B**).

Pues bien, el test con x y el test con y por separado, son tests *t distintos* del test T^2 empleado con (x, y), equivalente a una F. Tales tests no tienen por qué dar resultados compatibles. Las probabilidades de las regiones de rechazo son distintas. Además, la potencia del test con (x, y) es superior, puesto que la probabilidad de la región **F** es mayor que las probabilidades sumadas de las regiones **A** y **B**.

Para otras explicaciones de esta paradoja, véase Cramer (1975).

5. Correlaciones que no alcanzan el valor uno

El coeficiente de correlación ρ entre dos variable X, Y es un valor que oscila entre -1 y $+1$. Pero si las variables siguen distribuciones de distinta familia no pueden alcanzar tales valores. Se demuestra que si las funciones de distribución son F y G, ambas funciones contínuas, y las variables están estandarizadas, entonces las correlaciones mínima y máxima son

$$\rho^- = \int_0^1 F^{-1}(t)G^{-1}(1-t)dt \quad \text{y}$$
$$\rho^+ = \int_0^1 F^{-1}(t)G^{-1}(t)dt.$$

Cualquiera que sea la distribución de probabilidad conjunta de (X, Y) proporcionando un coeficiente de correlación $\rho(X, Y)$, se verifica

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

Es muy fácil ver que resulta imposible que $\rho(X, Y)$ alcance el valor 1 si X es uniforme e Y

es exponencial. Pues si así fuera, existiría una combinación lineal entre ambas, $Y = aX + b$, y por lo tanto Y seguiría también una distribución uniforme, cambiando sólo la media y la varianza, y no una exponencial.

6. Solución a la paradoja del coeficiente de correlación

Se suponía que X, Y eran dos variables aleatorias definidas sobre la misma población, con covarianza σ_{XY}, variancias finitas σ_X^2, σ_Y^2 y coeficiente de correlación de Pearson $\rho = \sigma_{XY}/(\sigma_X \sigma_Y)$. Seguidamente se tomaban $X_1, \ldots, X_n$ 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. Suponiendo $\text{cov}(X_i, Y) = \sigma_{XY}$, se probaba que la correlación entre la media $\overline{X}_n$ y la variable Y es $\sqrt{n}\rho$. Luego, para n suficientemente grande, el coeficiente de correlación entre la media $\overline{X}_n$ e Y puede ser "mayor" que 1. Veamos que esto es imposible.

Primera explicación (graciosa): Suponiendo que la correlación entre la estatura Y del padre con la X del hijo es $\rho = 0,5$, ningún padre puede tener más de 4 hijos varones (con distintas mujeres) para evitar que $\sqrt{n}0,5$ supere el valor 1. De hecho, no se conoce ningún caso con tantos hijos varones nacidos de distinta mujer.

Segunda explicación (seria): Si X e Y están correlacionadas, no es posible tomar una muestra $X_1, \ldots, X_n$ de valores independientes de X. La independencia de la muestra es incompatible con que esté correlacionada con Y. Se desprende de $\sqrt{n}\rho \leq 1$ que los valores x son necesariamente dependientes. Así, las estaturas de los hijos que comparten un mismo padre están necesariamente correlacionadas.

Esta paradoja nos advierte de que, en ciertas situaciones, no se puede tomar alegremente una muestra de tamaño n de valores independientes.

La anécdota: se propuso esta paradoja a un destacado probabilista puro de una universidad británica, pero fue incapaz de resolverla. En cambio, otros estadísticos y probabilistas (aunque no todos), más familiarizados con la estadística, la resolvieron rápidamente.

7. Solución a la predicción racista

Vamos a resolver la paradoja con un ejemplo. Supongamos que todas las medias de las variables I y H valen 100 en los grupos B y N, excepto la media de I que es 90 en el grupo N. Todas las desviaciones típicas valen 12, y los coeficientes de correlación, tanto en B como en N son $r = 0,7$. Se argumentaba que si un individuo B (blanco) posee el mismo I (coeficiente de inteligencia) que otro individuo N (negro), la predicción de H sería superior. Sin embargo, la predicción es incorrecta, sucediendo justo al revés. En efecto, las rectas de regresión son

$$\text{B:} \quad H = 100 + 0,7(I - 100)$$
$$\text{N:} \quad H = 100 + 0,7(I - 90)$$

Entonces si un B y un N puntuan igual $I = 110$, las predicciones de H son

$$H = 100 + 0,7(110 - 100) = 107 \quad \text{(individuo B)},$$
$$H = 100 + 0,7(110 - 90) = 114 \quad \text{(individuo N)}.$$

Así, para un mismo nivel de inteligencia 110, la predicción para la habilidad H es superior en N que en B. La figura 2 ilustra esta paradoja. La recta B es paralela y está situada a la derecha de N por ser la media de I más alta. Sin embargo, a un mismo valor de I le corresponde un valor de H más alto en el grupo N.

Este error de interpretación aparece en la controvertida obra *The Bell Curve*, de Herrstein y Murray. Véase Kaplan (1997) y Cuadras (2003).

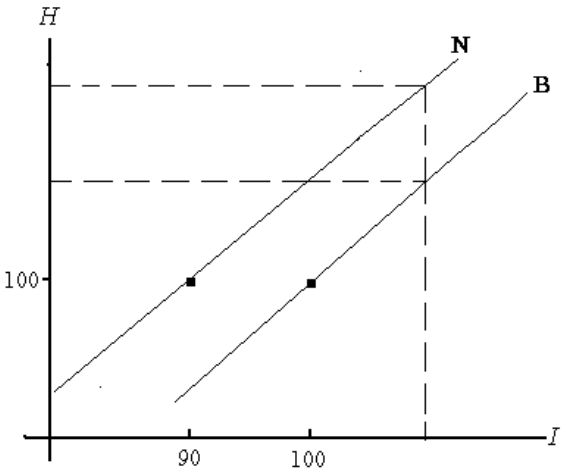

Figura 2. Una media mayor para I (inteligencia) en el grupo B no implica, comparando con otro grupo N, una predicción mejor para otra característica correlacionada H cuando I ha tomado el mismo valor en B y N.

8. Correlaciones simples aumentando y la múltiple disminuyendo

En efecto, puede suceder que aumentando las correlaciones simples disminuya la correlación múltiple. Esta aparente anomalía para variables equicorrelacionadas, fue primeramente observada por Tiit (1984). Vamos a formular una explicación en el caso general.

Supongamos que la variable respuesta Y correlaciona con $X_1, \ldots, X_k$, según el vector $\mathbf{r}$, siendo $\mathbf{R}$ la matriz de correlaciones entre las x's. El coeficiente de correlación múltiple (al cuadrado) es

$$R^2 = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r}.$$

Consideremos la descomposición espectral de $\mathbf{R}$ y de su inversa $\mathbf{R}^{-1}$

$$\mathbf{R} = \sum_{i=1}^{k} \lambda_i \mathbf{u}_i \mathbf{u}_i', \qquad \mathbf{R}^{-1} = \sum_{i=1}^{k} \frac{1}{\lambda_i}_{i} \mathbf{u}_i \mathbf{u}_i',$$

siendo $\lambda_1 > \cdots > \lambda_k$ los valores propios y $\mathbf{u}_1, \ldots, \mathbf{u}_k$ los vectores propios ortonormales. Entonces la correlación múltiple (al cuadrado) es

$$R^2 = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} = \sum_{i=1}^{k} \frac{1}{\lambda_i} (\mathbf{r}'\mathbf{u}_i)^2.$$

Como la suma de los valores propios es k, el primer valor propio es mayor que 1 y el último es menor que 1. Resulta entonces que si $\mathbf{r}$ sigue esencialmente la dirección de $\mathbf{u}_k$, entonces $\mathbf{r}'\mathbf{u}_k$ puede tener un peso importante en R^2.

Para los dos ejemplos propuestos

$$\mathbf{r}_1 = \begin{pmatrix} 0,6 \\ 0,5 \\ 0,4 \\ 0,3 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & 0,3 & 0,4 & 0,5 \\ 0,3 & 1 & 0,5 & 0,4 \\ 0,4 & 0,5 & 1 & 0,3 \\ 0,5 & 0,4 & 0,3 & 1 \end{pmatrix}$$

$$\mathbf{r}_2 = \begin{pmatrix} 0,6 \\ 0,5 \\ 0,1 \\ 0,1 \end{pmatrix}$$

los productos escalares normalizados son $\mathbf{r}_1 \cdot \mathbf{u}_4 = 0,2157 < \mathbf{r}_2 \cdot \mathbf{u}_4 = 0,5669$. Es decir, $\mathbf{r}_2$ forma un ángulo con $\mathbf{u}_4$ menor que $\mathbf{r}_1$. Entonces las correlaciones múltiples (al cuadrado) son

$$R_1^2 = 0,4848 = 0,3682 + 0,0000 + 0,0167 + 0,1000,$$
$$R_2^2 = 0,7056 = 0,1920 + 0,0031 + 0,0042 + 0,5052.$$

Tenemos pues que

$$\mathbf{r}_1'\mathbf{r}_1 = 0,86 > \mathbf{r}_2'\mathbf{r}_2 = 0,63 \quad \text{pero} \quad R_1^2 < R_2^2.$$

En otras palabras, cuando la dirección de $\mathbf{r}$ con las correlaciones simples, es próxima a la de un vector propio de $\mathbf{R}$ asociado a un valor propio menor que 1, la correlación múltiple puede tomar un valor sorprendentemente alto. Para más detalles, véase Cuadras (1995).

En realidad, la variable respuesta Y estaría demasiado correlacionada con las últimas componentes principales obtenidas a partir de $\mathbf{R}$, lo que provoca una cierta distorsión. Es decir, como se comenta en la sección siguiente, se confirma la importancia de la primera componente principal en el comportamiento de las variables explicativas. Se puede también argumentar que si las variables X están positivamente correlacionadas, y la respuesta Y correlaciona positivamente con una, debería correlacionar también positivamente con las demás. Si así ocurriera, lo que es bastante razonable, Y apenas correlacionaría con las últimas componentes principales.

La anécdota: esta peculiaridad en regresión (aumentando las correlaciones simples disminuye la múltiple) se presentó en un congreso internacional en 1994. Pero la posibilidad de que Y correlacionara de manera distinta con las variables explicativas, fue negada categóricamente por un destacado estadístico de Stanford, provocando una acalorada discusión entre partidarios y detractores de los argumentos aquí presentados.

9. Explicación a una desigualdad de la correlación múltiple

Con las mismas notaciones que en la sección anterior, vamos a estudiar la sorprendente desigualdad

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

que prueba que variables correlacionadas no son siempre redundantes, y que a veces mantienen una estructura de dependencia que es más difícil de interpretar de lo que parece.

La desigualdad puede expresarse como

$$\mathbf{r}'\mathbf{R}^{-1}\mathbf{r} - \mathbf{r}'\mathbf{r} = \sum_{i=1}^{k} \frac{1-\lambda_i}{\lambda_i} (\mathbf{r}'\mathbf{u}_i)^2 > 0.$$

De nuevo vemos que $(\mathbf{r}'\mathbf{u}_i)^2$ influye mucho si λ_i es un valor propio menor que 1 y $\mathbf{r}$ sigue esencialmen-

te la dirección de $\mathbf{u}_i$, en especial la dirección de $\mathbf{u}_k$. Esto es precisamente lo que ocurre en el ejemplo anterior con la segunda variable respuesta

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

Se puede probar que si $\mathbf{r}$ sigue esencialmente la dirección de $\mathbf{u}_k$, entonces la respuesta Y está muy correlacionada con la última componente principal. Más exactamente, la desigualdad anterior equivale a

$$\sum_{i=1}^{k} r_{z_i}^2 (1 - \lambda_i) > 0,$$

donde r_{z_i} es la correlación simple entre Y y la componente principal Z_i. Entonces la influencia de r_{z_i} es relevante si $1 - \lambda_i > 0$, como ocurre con la última componente principal.

Las componentes principales con varianza pequeña, en especial la última, indican las direcciones extrañas del conjunto de variables explicativas. En ciertas aplicaciones se interpretan como direcciones de "error". Podemos afirmar que se presenta la desigualdad objeto de este estudio si la variable respuesta sigue esencialmente la misma dirección que las últimas componentes principales, una situación no deseable pero que puede ocurrir con datos reales. Véase Cuadras (1993, 1998) para más detalles técnicos y ejemplos.

10. ¿Mahalanobis mayor que Pearson?

La desigualdad $M > K$, donde $M = (\overline{\mathbf{x}} - \overline{\mathbf{y}})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ es la distancia de Mahalanobis entre dos poblaciones y $K = (\overline{\mathbf{x}} - \overline{\mathbf{y}})'[\mathrm{diag}(\mathbf{S})]^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ es la distancia de K. Pearson, se presenta cuando $(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ sigue esencialmente la dirección de una componente principal con varianza pequeña.

Vamos a concretar la desigualdad para el caso de la última componente principal. Como la suma de los valores propios es la traza de $\mathbf{S}$, podemos suponer que el menor valor propio del vector propio u_k verifica

$$\lambda_k < s_i^2, \quad i = 1, \ldots, k.$$

Supongamos, por ejemplo, que $(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = \alpha \mathbf{u}_k$. Entonces $\mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) = (\overline{\mathbf{x}} - \overline{\mathbf{y}})/\lambda_k$ y la distancia de Mahalanobis verifica

$$\begin{aligned} M = \quad & (\overline{\mathbf{x}} - \overline{\mathbf{y}})'(\overline{\mathbf{x}} - \overline{\mathbf{y}})/\lambda_k \\ & (\tfrac{x_1 - y_1}{\lambda_k})^2 + \cdots + (\tfrac{x_k - y_k}{\lambda_k})^2 > \\ & (\tfrac{x_1 - y_1}{s_1})^2 + \cdots + (\tfrac{x_k - y_k}{s_k})^2 = K \end{aligned}$$

Un argumento más complicado pero similar, permitiría estudiar la desigualdad $M > K$ cuando $(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ sigue esencialmente la dirección de las últimas componentes principales. Como en el caso de la regresión múltiple, la interpretación de $M > K$ es que las matrices de datos $\mathbf{X}, \mathbf{Y}$ siguen en cada población, la dirección determinada por las primeras componentes principales. Sin embargo, el vector que une la medias sigue una dirección básicamente ortogonal. En otras palabras, como muestra la figura 3, las medias de las poblaciones no siguen la misma dirección que los datos en cada población (paradoja de Simpson).

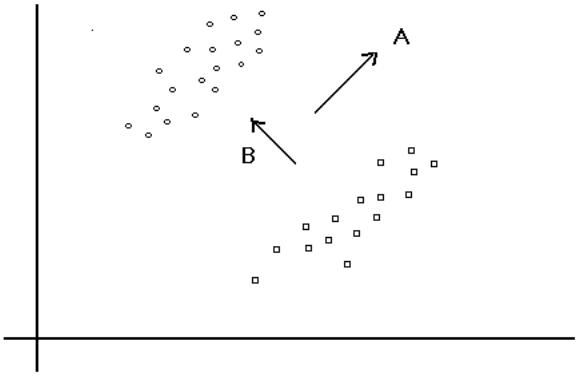

Figura 3. Los datos para cada una de las dos poblaciones siguen la dirección A (primera componente principal), pero las medias de las poblaciones siguen la dirección ortogonal B (segunda componente principal).

11. Por qué los momentos no siempre caracterizan

Hay un dicho citado por Francisco de Quevedo que dice: "Dime con quién fueres y direte cuál eres". Trasladado a las distribuciones de variables estadísticas, podríamos afirmar: "Dime qué momentos tienes y te diré cómo te distribuyes". Pero este dicho puede fallar, pues hay distribuciones que no están caracterizadas únicamente por sus momentos. Un ejemplo importante es la distribución log-normal con densidad

$$f(x) = (2\pi)^{-1/2} \frac{1}{x} \exp[-\frac{1}{2}(\log x)^2] \quad \text{para } x > 0.$$

Dos condiciones para que la sucesión $\alpha_n = E(X^n)$ de los momentos de todos los órdenes no caractericen la distribución de la variable son:

$\int_{-\infty}^{+\infty} \frac{-\ln f(x)}{1+x^2} dx < \infty$ si el soporte de f es R,

$\int_{-\infty}^{+\infty} \frac{-\ln f(x^2)}{1+x^2} dx < \infty$ si el soporte de f es R_+.

La no caracterización significa que existen dos distribuciones distintas que tienen los mismos momentos. Una explicación sencilla e intuitiva consiste en tener en cuenta que los momentos son valores esperados $\int_R x^n f(x)dx$, y en consecuencia pueden proporcionar el mismo valor si perturbamos $f(x)$ de modo que las integrales (que son cantidades medias) se compensen. Para profundizar más en este tema, véase Stoyanov (1997, p. 101).

12. Función generatriz que no distingue

Si bien los momentos podrían no distinguir, es en cambio cierto 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. No obstante existen distribuciones distintas para las cuales apenas se distinguen (numérica y gráficamente) las funciones generatrices. Por ejemplo:

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

La explicación transcurre por el mismo camino que los momentos comunes en distribuciones distintas. Al ser $M_X(t)$ un valor medio que depende de t, para ciertas distribuciones, como las mencionadas, los valores medios se compensan y dan lugar a funciones muy parecidas.

La situación cambia radicalmente si tomamos la función característica

$$\varphi_X(t) = E(e^{itX}) = \int_a^b e^{ixt} dF(x).$$

Como prueba Waller (1995) el uso de $\varphi_X(t)$ da lugar a funciones (de variable real a valores complejos) que pueden ser bastante distintas, debido a la presencia de la parte imaginaria. En el caso que nos ocupa, las funciones características son

$$\varphi_X(t)=e^{\frac{(it)^2}{2}},\ \varphi_Y(t)=e^{(it)^2/2+\log(1+\frac{1}{2}e^{-2\pi^2}\sin(2\pi it))}.$$

Su representación da lugar a gràficos iguales para la parte real, pero diferentes para la parte imaginaria y por supuesto distinguibles, como muestra la figura 4.

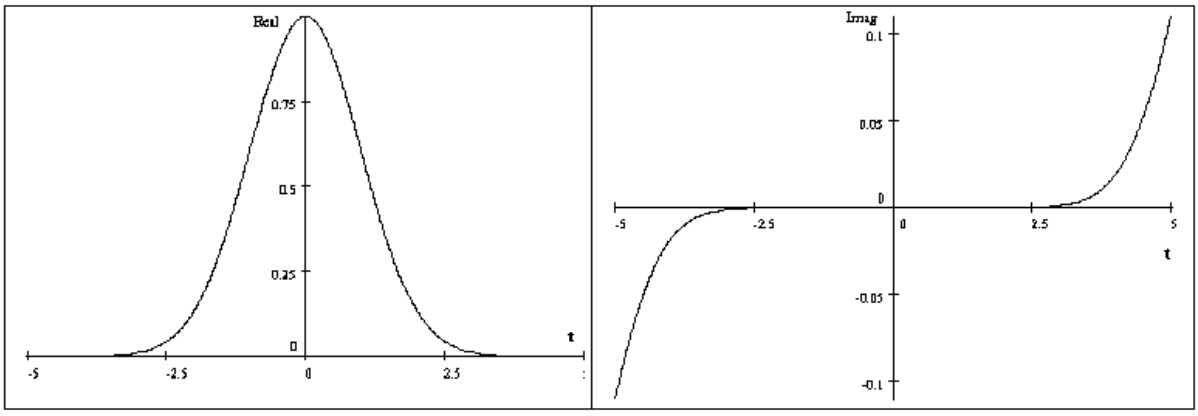

Figura 4. La parte real (izquierda) de las funciones características $\varphi_X(t)$, $\varphi_Y(t)$ es indistinguible. Sin embargo la parte imaginaria (derecha) vale 0 para $\varphi_X(t)$, y es distinta de 0 para $\varphi_Y(t)$ si $|t|$>2,5, pudiéndose distinguir una de otra.

En definitiva, se puede afirmar que la función generatriz (basada en la transformación de Laplace) es interesante para encontrar momentos y probar propiedades de ciertas distribuciones, pero es poco útil para distinguirlas numéricamente. En contraste, la función característica (basada en la transformación de Fourier) permite comparaciones numéricas mucho más eficientes.

13. La ley de los grandes números no falla

Se denunciaba que si X es una variable aleatoria distribuida Poisson con media $\lambda = 1$, entonces la media $\overline{X}_n$ de n valores independientes verifica $\overline{X}_n \xrightarrow{P} 1$, o mejor dicho :

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

Sin embargo, mostrábamos que

$$\lim_{n\to\infty} P(\overline{X}_n = 1) = \frac{e^{-n}n^n}{n!} = 0.$$

Es decir, a pesar de que $\overline{X}_n$ converge casi seguramente a 1, $\overline{X}_n$ no puede alcanzar *exactamente* el valor 1 si hacemos tender n a infinito.

Aunque sorprenda a primera vista, la imposibilidad de alcanzar $\overline{X}_n$ el valor medio teórico 1 no contradice la famosa ley de los grandes números. En realidad ocurre que

$$\lim_{n\to\infty} P(|\overline{X}_n - 1| > \epsilon) = 0,$$

por pequeño que sea $\varepsilon > 0$. Es decir, $\overline{X}_n$ tomará valores en un entorno $(1-\varepsilon, 1+\varepsilon)$ con certeza absoluta. También podemos interpretar que la distribución de $\overline{X}_n$, a medida que n crece, se aproxima a la normal, y es bien sabido que para una distribución contínua la probabilidad de que tome exactamente un valor concreto (conjunto de medida nula) es igual a cero.

14. El teorema central del límite no falla

Contradecíamos el famoso teorema central del límite tomando $X_1, \ldots, X_{100}$ Poisson independientes con parámetro $\lambda = 0,01$ y obteniendo la suma

$$X = X_1 + \cdots + X_{100},$$

que se distribuye según una Poisson con media $\lambda = 1$. Por lo tanto la distribución de X es demasiado distinta de la normal.

Pero el teorema no se contradice, tratándose de un simple truco, propio de un estadístico veterano e intrigante. En efecto, podemos sumar mil Poissones y las que queramos y "contradecir" el teorema, con tal de tomar $\lambda = 0,001$ o cualquier λ suficientemente pequeño. En realidad estamos sumando muchas variables con varianza muy pequeña, variables aleatorias que son "casi" constantes, de modo que la suma da lugar a una variable con varianza 1.

Este aparente incumplimiento también ocurre con la distribución binomial $B(n,p)$, cuya variable es suma de n Bernoullis independientes. Pues si n es muy grande y p muy pequeño, la distribución $B(n,p)$ es aproximadamente Poisson, con $\lambda = np$. Por ejemplo, es Poisson $\lambda = 1$ si $n = 1000$ y $p = 0,001$. Tampoco se contradice el teorema central del límite, alertando estos dos ejemplos de que, bajo ciertas circunstancias, la suma de muchas variables independientes puede proporcionar una distribución alejada de la normal.

15. Por qué un test de multinormalidad resulta poco efectivo

Basándose en un teorema debido a H. Crámer, se proponía aceptar la normalidad multivariante de $X_1, \ldots, X_k$ tomando la suma $Z = Y_1 + \cdots + Y_k$, donde $Y_1, \ldots, Y_k$ son las componentes principales extraídas de una matriz de datos $\mathbf{X}$ de orden $n \times k$, con n grande. La normalidad univariante de Z debería garantizar la multinormalidad de $X_1, \ldots, X_k$.

Desde un punto de vista probabilístico, el resultado es correcto. Z es normal si la distribución de $X_1, \ldots, X_k$ es normal multivariante. Pero... una cosa es la probabilidad, basada en modelos matemáticos, a menudo descritos mediante funciones muy bonitas, y otra distinta la estadística, siempre basada en datos reales producto de la observación experimental. En efecto, si aplicamos este "test", al que llamaremos CC (Crámer-Cuadras) detectaremos fácilmente que una muestra $\mathbf{X}$ sigue la distribución multinormal cuando ésta es la verdadera distribución de las filas de $\mathbf{X}$. CC funciona bien cuando el modelo multinormal es el verdadero. Pero CC no pasará a la posteridad, ni merecerá aparecer en el buscador Google porque si $\mathbf{X}$ no es multinormal, CC también detectará multinormalidad. Es decir, el "test" CC en la inmensa mayoría de los casos detectará multinormalidad, tanto si los datos siguen la normal multivariante como si no.

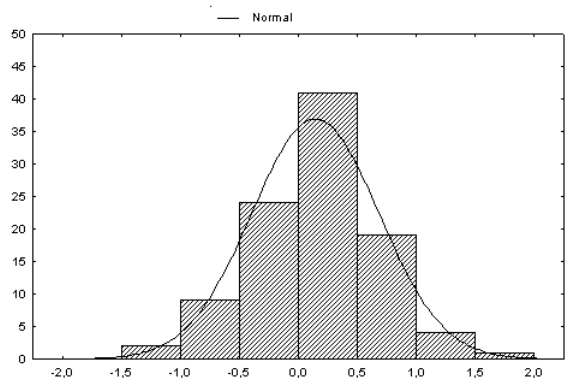

Figura 5. El test de multinormalidad basado en la suma de las componentes principales no permite distinguir (como en este caso de variables que son potencias de uniformes) si los datos proceden de una distribución normal multivariante o no.

¿Por qué? Al ser Z una suma de k componentes principales, que son variables incorrelacionadas, aparecerá un efecto debido al teorema central del límite, y la distribución de Z, de la que sólo dispondremos de una muestra de tamaño n, se parecerá demasiado a la normal, hasta el punto de que un test de normalidad univariante nos inducirá a aceptar la hipótesis nula.

Por ejemplo, generando una tabla con $n = 100$, $k = 4$, datos uniformes $(0, 1)$ e independientes, y transformando cada variable X_i elevándola a la potencia i, es evidente que la distribución conjunta no es multinormal. Sin embargo, la variable Z se ajusta bastante bien a la normal (test de Kolmogorov-Smirnov= 0,056, con $p > 0,20$ en la tabla de Lillefors), véase la figura 5. El "test" CC indicaría erróneamente que la tabla se ajusta a la normal multivariante.

La anécdota: este cándido planteamiento fue el primer intento de trabajo de investigación de un estadístico joven y novato, que interpretó al pie de la letra una propiedad probabilística de la distribución normal.

Referencias

[1] Cramer, E. M. (1975). The relation between Rao's paradox in discriminant analysis and regression analysis. *Multivariate Behavioral Research*, **10**, 99-107.

[2] Cuadras, C. M. (1993). Interpreting an inequality in multiple regression. *The American Statistician,* **47**, 256-258.

[3] Cuadras, C. M. (1995). Increasing the correlations with the response variable may not increase the coefficient of determination: a PCA interpretion. In: *Multivariate Statistics and Matrices in Statistics*, pp. 75-83. (E. M. Tiit, T. Kollo and H. Niemi, eds.), VSP/TEV, Utrecht.

[4] Cuadras, C. M. (1998). Some cautionary notes on the use of principal components regression. (Revisited). *The American Statistician*, **52**, p. 371.

[5] Cuadras, C. M., Fortiana, J. (2000). The Importance of Geometry in Multivariate Analysis and some Applications. In: *Statistics for the 21st Century*, pp. 93-108, (C.R. Rao and G. Szekely, eds.), Marcel Dekker, New York.

[6] Cuadras, C. M. (2003). *Report. Una narració científica*. EUB, Barcelona.

[7] Cuadras, C. M. (2007). Algunas paradojas y curiosidades de la estadística. *Boletín de la SEIO*, **23**(1), 24-29.

[8] Kaplan, J. (1997). A statistical error in *The Bell Curve*. *Chance*, **10**, 20-21.

[9] Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. Wiley and Sons, New York.

[10] Rao, C. R. (1952). *Advanced Stastistical Methods in Biometric Research*. John Wiley and Sons, New York.

[11] Stoyanov, J. (1997). *Counterexamples in Probability*. John Wiley and Sons, Chichester, New York.

[12] Székely, G. (1986). *Paradoxes in Probability Theory and Mathematical Statistics*. P. Reidel Pub. Co., Dordrecht, Boston.

[13] Tiit, E. M. (1984). Formal computations of regression parameters. In: *Proceedings Sixth Symposium COMPSTAT* 1984, pp. 497-502. (T. Havraneek, ed.), Physica-Verlag, Vienna.

[14] Waller, L. A. (1995). Does the characteristic function numerically distinguish distributions? *The American Statistician,* **49**, 150-152.