ACERCA DE LAS MEDIDAS DE ASOCIACIÓN EN INVESTIGACIÓN SOCIAL: UN VIEJO PROBLEMA QUE CONVIENE NO OLVIDAR

Luis Alfonso Camarero Rioja

Ha pasado ya más de un siglo desde que Sir Karl Pearson formulara el coeficiente de correlación. Exitoso y afamado estadístico que ha llegado a fundar una métrica de interacción de las variables con aplicaciones evidentes para todos los campos del conocimiento. De forma inmediata en el tiempo y dentro del propio entorno académico de Pearson se realizaron los primeros planteamientos para transportar el célebre «r» desde las variables gaussianas continuas hasta las variables nominales que dan lugar a las tablas de contingencia.

La tabla de contingencia, el simple conteo de frecuencias por pares de atributos, es sin duda la materia prima de la sociología cuantitativa. Sin embargo, cien años más tarde desde el «descubrimiento» de Pearson y después de un tortuoso camino de búsqueda de medidas alternativas a «r» que fueran válidas para las tablas de contingencia —itinerario que es el argumento de estas líneas—, los sociólogos siguen estableciendo e interpretando las relaciones entre variables casi exclusivamente desde el recurso de la lectura de porcentajes. La euforia que, en algunos momentos, han despertado ciertos coeficientes de asociación o modelos, no ha pasado de ser una moda, sin que ninguno de las decenas de índices propuestos haya alcanzado la sanción de la popularidad, ni superado los distintos problemas teóricos y técnicos.

En las páginas siguientes se realiza un recorrido por las distintas elaboraciones y propuestas que se han desarrollado para atacar el problema de la asociación entre variables cualitativas. El recorrido no es

cronológico, y el conocimiento, como sabemos, no es lineal, sino lógico, y busca poner en evidencia los supuestos, la mayor parte de las veces implícitos, que configuran distintas visiones de entender las relaciones entre modalidades y categorías.

El paso del tiempo nos permite ahora observar que en buena medida la dificultad para alcanzar una solución satisfactoria ha provenido de un mal planteamiento del problema. La fuerza de las corrientes positivistas a principios del siglo XX era tal que se cayó en la tentación de la analogía y así se comenzó a pensar las tablas de contingencia, entonces llamadas simplemente de atributos, de forma similar a una función. Desde esta formulación analógica se pensaba que las operaciones propias de la métrica intervalar deberían tener sus equivalentes en una supuesta «métrica nominal»[1]. La única precisión que el tiempo ha ido sancionando es el uso del término correlación para medir la relación entre variables continuas, dejando el genérico de asociación para las nominales.

El índice de correlación de Pearson es un índice que muestra la dirección y fuerza de la asociación dentro de un intervalo de valores cerrado, es insensible a las transformaciones de escala, es escalado a través de la función cuadrática o coeficiente de determinación y resulta muy intuitivo de interpretar, dado el rango de valores acotado que alcanza. Este índice ha sido el canon a copiar por quienes se han ocupado de las tablas de contingencia.

EL ORIGEN DE UN PROBLEMA MAL PLANTEADO: LA CONTINGENCIA COMO FUNCIÓN

El propio Pearson extendió, a través del coeficiente de correlación punto biserial, el desarrollo del coeficiente de correlación al caso en que la variable dependiente era binomial en vez de normal. Su discípulo Yule propuso en 1900 el coeficiente Q, llamado así en honor a Quetelet, como un coeficiente para el análisis de las tablas de contingen-

[1] Pero de hecho no existe una métrica nominal. No será, como se verá más adelante, hasta los años setenta cuando BENZECRI dote a las tablas de contingencia de una estructura topológica que permita fundar un espacio métrico.

cia más simples, las dicotómicas o tablas 2x2. Pearson fue muy crítico con la propuesta de Yule por dos motivos. El primero, la Q de Yule no tenía en cuenta la distribución de probabilidad de origen, Pearson se preguntaba sobre el carácter no normal de la distribución. Con el tiempo, ésta llegará a ser una crítica menor en la medida en que el coeficiente de correlación se ha extendido también a otras distribuciones no normales. En segundo lugar, Pearson criticaba la interpretación y significado del coeficiente Q, y de forma sarcástica le dirá a Yule que por qué no utiliza Q elevado a la tres o la quinta potencia[2]. El hecho es que el coeficiente «r» de Pearson tiene una interpretación inmediata a través del coeficiente de determinación r^2 que señala la parte de la varianza compartida que es explicada por la variable independiente, pero no así el coeficiente Q.

El coeficiente Q puede analíticamente describirse como B=QA, siendo B y A dos variables nominales dicotómicas. Es el equivalente a la ecuación Y=rX, siendo Y e X dos variables normales y normalizadas. (En ambas ecuaciones se ha omitido el término de error). En definitiva Yule lo que hace es establecer una función entre dos variables de atributos B=f(A), función que es lineal y así B=kA, siendo la constante k, el coeficiente Q. Por tanto B=QA. El problema estriba en que la relación entre A y B no es una función.

Yule tardará más de una década en solucionar el problema y contestar a su maestro. En 1912 Yule, propondrá el coeficiente de Coligación para tablas 2x2 que puede ser interpretado como un coeficiente de asociación entre dos variables binomiales que toman valores 0 y 1. Sencillamente Yule había hecho caso a su maestro, había tenido en cuenta la distribución de las variables y así había transportado el «r» a un caso específico de variables no normales. Abandonaba así su impulso original e ingenuo de considerar de forma genérica a las tablas de contingencia como funciones[3].

---

[2] La referencia a potencias impares era para que no variara el signo obtenido.

[3] En muchos manuales al uso suele denominarse Q de Yule al propio coeficiente de Coligación.

RELACIÓN, APLICACIÓN Y FUNCIÓN

Antes de seguir con esta exposición conviene expresar desde una postura de formalismo matemático la naturaleza de las tablas de contingencia, planteamiento éste que tardó en comprenderse y hacerse de forma explícita y de cuya ausencia se derivan buena parte de las insatisfacciones, caminos sin salida y especialmente las malas interpretaciones que se han dado a las diferentes propuestas.

El término de relación es un concepto muy amplio. Matemáticamente se establece una relación en cuanto que entre los elementos de dos conjuntos o del mismo conjunto pueden formarse pares entre ellos o de forma genérica tuplas. La relación es una forma de agrupamiento entre elementos. Esta operación cuando se establece entre clases de equivalencia, también llamadas fibras, se denomina producto cartesiano.

Un tipo especial de relación es la aplicación. La aplicación relaciona a dos conjuntos origen e imagen, siempre que todos los elementos del conjunto origen tengan una y sólo una imagen. Las tablas de contingencia no cumplen esta propiedad, no son aplicaciones. Ciertamente en una tabla todas las categorías de una variable están relacionadas con todas las de la otra variable[4].

Otro nombre que recibe la aplicación es el de función. Los valores que toma una variable son función de otra. Por ejemplo el consumo de gasolina de un automóvil es función del número de kilómetros recorrido. El coeficiente de correlación mide la intensidad de relación que produce una función lineal, señalándonos cuál es el efecto del incremento, medido en términos de varianza, de una variable sobre otra[5]. Este coeficiente puede hacerse más general a través del coeficiente «Eta» cuando se está ante funciones no lineales. El coeficiente de corre-

---

[4] Tan sólo en el caso en que nos encontráramos con una tabla cuadrada (kxk) en la que todas las frecuencias son cero a excepción de la diagonal, estaríamos ante el equivalente de una función. Sin embargo, esta relación sería la unión de k relaciones de identidad y no sólo una función identidad. Para aclarar esto véase más adelante la referencia al efecto Guttman.

[5] Como es sabido, «r» puede expresarse como la relación entre las desviaciones (siendo «a» la constante de la recta de regresión):

$$r = a \frac{\sigma_x}{\sigma_y}$$

lación puede aplicarse a casos en que la variable dependiente lo es de varias variables (r múltiple) pudiéndose además calcular y aislar el efecto parcial de cada una de ellas sobre la dependiente y sobre las otras independientes. El coeficiente de correlación mide la relación —tanto en su intensidad como en su dirección— que existe en una función o incluso entre un conjunto de funciones. Sin embargo su transporte a escalas de medida inferiores no es automático.

¿Es la tabla de contingencia una función? No, en absoluto. Es, simplemente, una relación definida mediante un producto cartesiano entre clases de equivalencia o fibras.

LA EVIDENCIA DE LA ASOCIACIÓN

Sin embargo, quien está acostumbrado a trabajar con tablas de contingencia observa relaciones entre los factores o variables e intuitivamente se pregunta por la intensidad de la asociación. Un ejemplo sencillo son las tablas de relación entre los estudios del padre y del hijo. Cuando son tablas cuadradas la inspección del «abultamiento» de la diagonal nos habla de la relación de la asociación y de ahí surge la primera idea, que es ingenua y errónea, y que consiste en considerar a la tabla de contingencia como una función y decir la variable B es función de la A, cuando lo que habría que decir es, simplemente, B está relacionada con A. Pero aunque no estemos ante una función, ¿podemos decir cuánto vale, medir o cuantificar esa relación? En el caso anterior, estudios de los padres y de los hijos, parece que la pregunta tiene sentido. Pero en este caso la tabla de contingencia contiene una relación de orden, ambas variables son ordinales aunque las presentemos como atributos. Imaginemos una tabla que relaciona nacionalidad y profesión de una población. En esta tabla también observamos algún tipo de asociación: en los países del Tercer Mundo, por ejemplo, encontraremos más campesinos y menos profesionales que en los del Primer Mundo, pero ¿cómo podemos cuantificar dicha relación?

Veamos a continuación las distintas respuestas y diferentes planteamientos intentados al respecto.

CRITERIO	FORMULACIÓN ANALÍTICA	COEFICIENTE/ AUTOR
- Función B=f(A)	B=QA	Q (Yule)
- Contingencia A*B=f(u)	A*B=f(a)f(b)+e	Ji-Cuadrado (Pearson)
	a_{ij}=f(a)f(b)+e	Residuos (Haberman)
	A*B=f(a)f(b)f(z)	Estructura latente (Lazarsfeld)
	A*B=f(a)W_af(b)W_bf(ab)W_{ab}	Loglinear (Goodman y Kruskal)
- Reducción Proporcional del Error A*B=(A*B)'+e	A=f(a/b)+e	Lambda (Goodman y Kruskal)
- Métrico	A*B:=d($a_i b_j$)	Correspondencias (Benzecri)

LA TABLA BIVARIABLE COMO CONTINGENCIA. LA CONQUISTA DE LA INDEPENDENCIA

Pearson también se preocupó por el tema e inmediatamente se puso a trabajar en las tablas de contingencia pero reconociendo su carácter de conjunto de frecuencias o de contingencias entre sucesos. Para ello utilizará el segundo de sus célebres coeficientes el Ji-cuadrado. En el planteamiento de Pearson subyace la consideración de la tabla como un conjunto imagen de una función de probabilidad. Es decir, la tabla es una configuración concreta determinada por una función.

Para medir la asociación lo que hace Pearson es construir un modelo teórico para cada tabla, modelo que se determina desde el supuesto de independencia absoluta a través de lo que se denominan frecuencias esperadas. Construye así una escala que tiene un origen y cuyo valor es 0. Posteriormente se compara la tabla de datos que se tiene con el supuesto de independencia. Esta operación, sin embargo, no está dotada de límite superior, ni desde luego hace referencia a la dirección de la asociación y es muy sensible al número de observaciones. Esta última objeción la resuelve el propio Pearson mediante el coeficiente Phi, coeficiente que teóricamente debería oscilar entre 0 y 1, pero que en la práctica, para determinadas composiciones de tablas, puede superar este acotamiento máximo.

A partir de los trabajos de Pearson se inaugura lo que podría llamarse la variante clásica de los coeficientes de asociación. Se propo-

nen distintos coeficientes construidos a partir de la medida de Ji-Cuadrado que buscan, mediante operaciones aritméticas, corregir los defectos del coeficiente Phi respecto al acotamiento superior. Así aparecen los coeficientes de Contingencia, Tschruprow y Cramer[6], si bien ninguno de ellos consigue su propósito de forma satisfactoria. El coeficiente de Cramer, llamado[7] V, es un valor Ji-Cuadrado estandarizado. El problema ahora es el de su interpretación. Los valores que alcanzan las medidas basadas en Ji-cuadrado son arbitrarios. De su análisis técnico se deduce su profunda sensibilidad a los marginales desequilibrados, pero su defecto más importante es que, aunque sea una medida acotada, no es escalada.

Lo que sí que ha permitido el Ji-cuadrado es construir contrastes de independencia estadística. Los tests de independencia se han desarrollado con gran éxito para casi todos los supuestos de configuraciones posibles, bajo cualquier hipótesis de partida o función generadora de las distribuciones marginales. Podemos así saber si dos variables nominales están o no asociadas, aunque no cuánto.

Desde el mero transporte del coeficiente «r» que planteaba Yule al coeficiente Phi de Pearson hay un cambio de óptica. El coeficiente Phi es una medida de inercia y por ello tiene interpretación[8]. Pearson lleva el debate mediante el uso del Ji-cuadrado a la comparación de modelos, pero fundamentalmente sitúa la polémica en la tabla de contingencia como configuración. La tabla es un resultado de la acción de dos funciones independientes. El uso del Ji-cuadrado evita un problema serio, el del conocimiento de las funciones de probabilidad que

---

[6] El propio PEARSON intentó superar el escollo del límite superior de Phi, mediante el coeficiente de contingencia; sin embargo, este coeficiente, acotado, que nunca puede ser mayor que la unidad, no permite comparar tablas cuando el número de casos es distinto. Un paso más adelante será el coeficiente de Tschruprow, que permite tener un límite superior acotado pero que varía con los grados de libertad. Este índice sólo alcanza el valor unidad cuando las tablas son cuadradas. El índice de Cramer resuelve esta dificultad, está acotado entre 0 y 1 y es independiente del tamaño de la tabla.

[7] El porqué de llamarse «V» es una cuestión no aclarada. Blalock, sin querer, es el mentor de esta medida, y cuando la expone como mejora sustantiva de los coeficientes «C» de contingencia y «T» de Tschruprow decide llamarla V, sin dar explicación alguna. Desde entonces queda bautizado como «V».

[8] El término «Inercia» proviene de los estudios de mecánica estadística. En la actualidad se usa para denominar a la varianza de distribuciones multivariables. Como se sabe,

$$I = \frac{\chi^2}{n}$$

están generando los marginales. Así podemos saber si dos variables son independientes sin necesidad de conocer el tipo de distribución que ha generado los datos. Este principio estará presente en la mayoría de los trabajos posteriores sobre contingencia, sin que ello obste para que supuestas otras funciones de probabilidad se puedan utilizar distintos test de independencia. (Test exacto de Fisher, McNemar...).

Analíticamente el planteamiento de Pearson lo podemos expresar haciendo que la tabla de contingencia como producto cartesiano A*B, sea el resultado del producto de dos funciones, desconocidas, de las variables nominales que intervienen.

$$A*B=f(A)f(B)$$

El producto f(A)f(B) es una relación desconocida excepto en el caso de que sea una relación aleatoria y considerándolo así obtenemos que A*B=f(A)f(B)+e, en donde el término de error nos está indicando precisamente el valor de la asociación.

¿QUÉ ES LA ASOCIACIÓN? LA REDUCCIÓN DE LA INCERTIDUMBRE

Pero el problema fundamental es conceptual. ¿Qué es la asociación? Cuando uno observa los manuales de estadística al uso, se sorprende de que no aparece una definición clara de la asociación. Valga como ejemplo la «no-definición» o definición tautológica que hace Somers para la *Enciclopedia Internacional de las Ciencias Sociales*, definición también contenida en la *Enciclopedia Internacional de Estadística* editada por Kruskal y Tanur:

> Cuando dos o más variables o atributos son observados para cada individuo de un grupo, la descripción estadística se basa frecuentemente en tablas de doble entrada que muestran el número de individuos que tiene cada combinación de valores de las variables. Además, se desea a menudo más brevedad y, en particular, se siente ordinariamente la necesidad de medidas (índices o coeficientes) que muestran en qué grado una variable está asociada a otra.[9]

[9] Artículo *Estadística Descriptiva: Asociación*, vol. IV, p. 418.

El Manual de BLALOCK es uno de los de mayor difusión para la generación de los sociólogos hoy «senior», y en vez de asociación habla de «fuerza de una relación». En el apar-

Del párrafo anterior destaca el recurso expresivo a la noción de «necesidad sentida». Es decir, se está diciendo que la asociación es algo que «desea» el investigador o analista pero que no pertenece a la propia estructura de los datos[10].

Debido a esa «necesidad», la asociación en tablas de contingencia ha sido pensada en términos de función. La asociación en una función está clara, es una medida del efecto que el incremento de la variable independiente tiene sobre la dependiente; ello, además, le dota de otras lecturas, la más profusa de las cuales es la interpretación predictiva. Así la asociación también se relaciona con el grado de éxito en la predicción de una variable a través de otra. Pero ¿cuál es la interpretación en el caso de las tablas de contingencia?. Si, por ejemplo, me encuentro con una tabla de relación entre religión y profesión, ¿qué sentido tiene hablar de un incremento de religión y su influencia sobre un incremento de profesión, o viceversa?

Una respuesta a esta pregunta sería: si conozco un atributo de un individuo, ¿en cuánto me ayuda este conocimiento para predecir con éxito otro atributo del mismo individuo? Así se plantea una noción específica del concepto de asociación en las tablas de contingencia, la reducción de la incertidumbre en un supuesto de predicción. Es decir, la asociación así interpretada responde a la pregunta: ¿en cuánto me ayuda conocer los valores de una variable para pronosticar los de otra? Desde estos planteamientos se desarrolla el denominado criterio de Reducción Proporcional del Error, conocido abreviadamente como RPE. Esta perspectiva ha desarrollado distintos coeficientes que, al contrario que en la variante clásica de medidas de asociación, son coeficientes asimétricos y que necesitan distinguir entre variable predictiva y predictora. Quizás el más célebre de estos coeficientes sea Lambda (que tiene en cuenta los valores modales) y Tau, de Goodman y Kruskal. Otra variante es el coeficiente de incertidumbre, que realiza lo mismo desde la teoría de la información. La línea de RPE ya había sido sugerida en los años cuarenta por Guttman, y son sobre todo los trabajos de Goodman y Kruskal los que desarrollan esta familia de coeficientes.

---

tado correspondiente se dedica a explicar las características ideales que debe tener un índice de fuerza de la relación, pero sin llegar en ningún momento a precisar formalmente este concepto.

[10] Más adelante se destacará cómo Hyman habla de «intimidad».

Estos coeficientes, en comparación con los de la variante clásica, comienzan a tener un sentido, en primer lugar porque se acercan y buscan una definición de asociación —se sabe lo que se mide—[11] y en segundo lugar porque están construidos como un incremento porcentual entre dos situaciones que lo dota de límites. La del desconocimiento absoluto con la del conocimiento que yo tengo.

En los métodos de reducción proporcional del error intervienen las frecuencias condicionadas. Analíticamente puede expresarse así:

$$A = f(A / B) + e$$

Es decir, la variable A es producto de una función condicional más un término independiente. El término «e», la parte no condicionada, o especificidad, tiene como un valor inversamente proporcional a la «cantidad de asociación». Uno de los problemas que plantea este método es la estimación de la función condicional. Por ello se recurre a comparar dos situaciones, de forma que el valor del coeficiente es el complemento del porcentaje que el término «e» ocupa en el valor de la igualdad.

El problema es que se construyen diversos coeficientes, porque ninguno acaba ofreciendo una solución definitiva. Unos son sancionados por los programas estadísticos y otros caen en el olvido, pero todos presentan grandes dificultades y ninguno llega a poder ser utilizado de forma general y universal. La principal dificultad la plantea el propio concepto de asociación perfecta. Existen varios tipos: perfecta estricta, que sólo es posible en tablas cuadradas, e implícitas de primer tipo, cuando todas las categorías de fila están asociadas con alguna categoría de fila, pero alguna categoría de columna se asocia con varias de la misma fila o viceversa, y de segundo tipo, cuando algunas categorías de columna se asocian con categorías de las mismas filas y viceversa. Estos coeficientes no reconocen la asociación perfecta. La asimetría de los coeficientes es otro problema, cuando no puede determinarse una

[11] Otra cuestión es la naturaleza epistemológica de dicha medida, que en este caso la asociación queda definida únicamente, y por tanto reducida a una medida de éxito en la pronosticación de valores.

variable causal de la relación. Por ejemplo, Lambda, que es una medida relativamente sencilla basada en los máximos, busca determinar el máximo de la variable dependiente. Infraestima la asociación en condiciones de marginales desequilibrados. En ciertas situaciones, lamentablemente muy frecuentes, las modas se encuentran en la misma fila o columna y, aun cuando hay asociación, el valor de Lambda es cercano a 0. Lambda, por ejemplo, no reconoce asociaciones perfectas implícitas y, lo que es peor, ofrece valores nulos en dichas situaciones. No tiene la propiedad de un único cero, es decir, varias configuraciones de la tabla, e incluso algunas de alta asociación pueden ofrecer un valor cero que debería estar reservado únicamente para el caso de independencia absoluta.

De hecho, cuando el analista maneja estas medidas se da cuenta de que tienen un comportamiento caótico, pequeñas variaciones en los datos son capaces de producir enormes variaciones en el valor del índice. Como mejora del coeficiente Lambda se propone «MR» (Rangos Múltiples) para alcanzar la propiedad de un único cero, y como mejora de éste, para dotarle de escala, «MP» (Probabilidades Múltiples)... En definitiva, el criterio de Reducción Proporcional del Error acaba desarrollando innumerables índices que son específicos para cada caso y, por tanto, pierden la generalidad que se le debe exigir a un coeficiente de asociación.

La mayoría de los analistas conocen estos índices no por planteamientos formales, sino a través de juegos didácticos. Veamos un ejemplo de cómo es estudiado Lambda y sus derivados:

> Supongamos jueces racionales que tratan de adivinar la columna de pertenencia de casos extraídos de una muestra cuando conocen las frecuencias de cada casilla para el conjunto de la muestra. La extracción de los casos es aleatoria y con reposición. Un juez «informado» dirá la fila de pertenencia del caso antes de realizar su pronóstico, mientras que un juez «ciego» no. Ambos jueces intentarán maximizar el número de adivinaciones correctas. El juez ciego siempre pronosticará la columna mediante la frecuencia total mayor, que será la mejor forma de acertar. El juez informado pronosticará la columna mediante la frecuencia mayor de la casilla perteneciente a la fila a la que sabe que pertenece el caso.
>
> (Darlington)

La ventaja estandarizada del juez informado sobre el ciego es el coeficiente Lambda. Este sistema de jueces ciegos e informados, que

en realidad son jugadores[12], se complica mediante sistemas de puntuación desiguales que penalizan los aciertos en las columnas más fáciles para generar otros índices derivados[13]. Pero el problema es cómo puede el analista interpretar con claridad un índice así expuesto... y cómo seguramente se habrá preguntado el lector: este ingenioso juego qué nos dice sobre nuestro problema, la asociación. La única conclusión posible es tautológica: Lambda y sus derivados tan sólo nos dicen que a más información mayor conocimiento.

¿CUÁLES SON LOS EFECTOS DE LA ASOCIACIÓN? LA HUIDA HACIA ADELANTE

Las dificultades para obtener medidas de asociación hacen reconsiderar el problema y se mira hacia otro lado. Si no podemos valorar la cantidad de asociación, ¿podemos al menos valorar los efectos de la misma?, ¿podemos determinar las contribuciones que tienen las modalidades y las variables entre sí?

En la tabla de contingencia los efectos de la asociación se pueden observar en los residuos, es decir, la diferencia entre las frecuencias esperadas y las frecuencias observadas. Este planteamiento sencillo es expuesto por Haberman mediante su análisis de los residuos. Se retoma la aplicación del Ji-cuadrado que había comenzado Pearson. Eso sí, Haberman provee al analista de una serie de utillajes técnicos que permiten determinar cuándo la asociación entre dos modalidades es estadísticamente significativa. Los residuos de Haberman indagan en las fuentes de la asociación. La extensión de esta lógica lleva a plantearse por modelos que estudien los efectos de las variables en el caso multivariante. Pero antes de exponer esta línea veamos los antecedentes de la misma en Lazarsfeld.

---

[12] El texto original dice «judge», siempre acompañado del adjetivo «rational». Pero el contexto hace que el «rational judge» sea un jugador o mejor un apostador. (Incluso me atrevería a decir que el «racional juez informado» puede parecer un jugador tramposo.)

[13] En la terminología expuesta por Darlington, los jueces racionales se convierten ahora en «perfectos» por equitativos.

LOS PRECEDENTES: EL ANÁLISIS DE LA ESTRUCTURA LATENTE

Los antecedentes en los modelos de efectos para datos nominales se encuentran en el análisis de estructura latente de Lazarsfeld elaborada a mediados de los sesenta. Lazarsfeld reconoce que su fuente de inspiración fue Guttman. Durante un tiempo los analistas forzaban los datos de manera que los atributos pudieran ser convertidos en alguna escala ordinal o dicotómica sobre los que mediante una codificación binaria 0-1 se pudiera utilizar alguna medida de asociación de funciones (correlación «r» o el coeficiente punto biserial, ambos de Pearson). Para casos multivariantes se utilizaban sumas de estos indicadores. Guttman a principios de los cuarenta señala que no es correcto utilizar atributos con elementos de métrica continua[14].

Lazarsfeld recoge de Guttman el interés por lo que entonces se llamaban «opiniones latentes» y de estas ideas elaborará el análisis de estructura latente. Entre las muchas herencias que la estadística y la sociología deben a Lazarsfeld está el hábito de control constante de las relaciones por terceras variables. Y de aquí surge la pregunta de Lazarsfeld: qué sucede cuando la relación que encontramos en una tabla de contingencia desaparece al ser controlada por las categorías de una tercera variable, es decir, cuando hay una tercera variable que explica la relación original. Lazarsfeld argumenta además que difícilmente el analista va a encontrar una variable que consiga determinar a otra, y, tanto si esto existe como si no, es debido a la acción de variables latentes no observadas y por tanto no medidas que bien inciden en la relación o que bien, por el contrario, la neutralizan.

El modelo de clases latentes es, en definitiva, un modelo teórico. De hecho, en cuanto que se utilizan variables que no son dicotómicas la estimación de los parámetros resulta compleja, incluso utilizando algoritmos de iteración, e imposible en muchos casos. Otro problema es la identificación e interpretación de las «clases latentes» que resultan.

En definitiva, las Estructuras Latentes consisten en un brillante desarrollo matemático sin aplicaciones útiles. Las «ecuaciones contables» de las tablas eran eso, «contables». En muchos casos se podía

---

[14] Guttman, como es conocido, propuso su escalograma, una escala de atributos jerárquica y acumulativa que produce una variable continua. De hecho, su procedimiento ha sido denominado análisis factorial cualitativo.

explicar la configuración de la tabla mediante una relación matemática aun sin saber su significado. Pero, ¿podrían existir otras relaciones que produjeran la misma configuración? Y ello, ¿podría ser cuantificable?

Analíticamente el modelo de estructura latente de Lazarsfeld puede descomponerse en la siguiente ecuación:

$$A*B=f(a)f(b)f(z)$$

Es decir, la tabla de contingencia es resultado del producto de las funciones de ambas variables más el producto de una o más variables ocultas o latentes.

LAS HOMOLOGÍAS CON EL ANÁLISIS DE LA COVARIANZA: MODELOS LOGARÍTMICO-LINEALES

Otro paso en la asociación de las variables nominales vendrá dado a finales de los setenta por autores como Goodman y Haberman, mediante los modelos logarítmico-lineales. La fuente de inspiración será el análisis de la covarianza. En el caso de una tabla de contingencia, la frecuencia de una casilla vendrá determinada por el producto o la suma en notación exponencial de los efectos de la variable columna, la variable fila y el efecto conjunto de ambas variables, además de un efecto fijo conocido como gran media. Este modelo completo se denomina saturado; el análisis consiste en contrastar distintas combinaciones de efectos respecto al modelo saturado y al modelo de independencia.

A la hora de la verdad, el uso de los análisis logarítmico-lineales se limita a buscar un modelo con pocos efectos, eliminar la interacción conjunta de un par o más variables y en algún caso el efecto de una variable completa. Esto no es siempre posible, y en la mayoría de las situaciones prácticas difícilmente se consigue un modelo simple que esté alejado del saturado. Otro problema es la adecuación del modelo resultante a la elaboración teórica de la tabla. Generalmente se sigue el criterio de jerarquía para poder interpretar correctamente el mode-

lo. En el caso de una tabla de contingencia bivariable el significado del efecto conjunto es oscuro. El efecto conjunto es definido como un resto respecto a la hipótesis de independencia.

Los modelos log-lineares guardan una semejanza enorme respecto a las estructuras latentes de Lazarsfeld.

$$A*B=f(a)W_af(b)W_bf(ab)W_{ab}$$

Como se ve, es un producto ponderado de las funciones de ambas variables más una tercera variable que es el efecto conjunto o interacción entre ambos factores.

Emparentado con el análisis logarítmico-lineal, pero próximos en su objeto a los modelos de regresión, se encuentran los modelos LOGIT, PROBIT, TOBIT, etc. En este caso la variable dependiente es cualitativa, los valores que toma se transforman a una función de probabilidad y se relaciona mediante regresión múltiple con las variables independientes. Si las variables independientes son cualitativas y se transforman en variables ficticias o *dummy*, estos modelos son aplicables al análisis de contingencia. Operando de esta forma se transforma la tabla de contingencia en una función. La solución a nuestro problema sería fácil, la asociación vendría medida por el correspondiente estadístico de bondad de ajuste del modelo. Pero aquí está el problema, la mala calidad de estos estadísticos. De hecho se llaman «pseudo r cuadrado» y aunque existen varias propuestas la clásica definición de pseudo R^2 es:

$$\frac{\chi^2}{\chi^2 + n}$$

Como se ve, este estadístico nunca alcanzará el valor unidad, aun cuando la asociación sea perfecta, y el lector se habrá dado cuenta de que tomando su raíz, es decir, «pseudo r», es lo mismo que el coeficiente C de Contingencia de Pearson.

HOMOLOGÍAS CON LAS RELACIONES CAUSALES: LA DIFERENCIA DE PROPORCIONES

Con el mismo propósito, pero desde una lógica de causalidad en los efectos, Davis en los setenta desarrolla su sistema de la diferencia de proporciones. El método de Davis construye una ecuación para cada casilla, siendo los coeficientes del efecto de la asociación, las diferencias de proporciones. Se trata de unas ecuaciones contables de la estructura de la tabla[15]. Para el caso de tablas de contingencia es otra forma de presentación y lectura de los porcentajes. Su interés estriba en su generalización multivariante, dada su posibilidad de representación mediante diagramas de flujo o de camino. En la práctica, sin embargo, cuando las variables son más de tres, y cuando alguna de las variables tiene cuatro o más modalidades, existen varios modelos alternativos, sin que sea posible su elección mediante el recurso a estadísticos de ajuste ni —lo que también es frecuente— mediante el recurso a modelos teóricos. De hecho su principal aplicación ha sido al estudio de procesos de cambio, es decir, cuando las variables pueden ordenarse temporalmente, como son el tiempo (series temporales) o las profesiones de padres e hijos (tablas de movilidad). En estos casos el analista no tiene grandes dudas sobre qué modalidad utilizar como base para calcular las diferencias.

Estos modelos, que a pesar de las imperfecciones apuntadas orientan al investigador en su digestión y comprensión de los datos, poco aportan al tema de la asociación. En el caso más sencillo, el de la tabla bivariable, permiten como mucho ordenar por su importancia los efectos que los factores tienen sobre las casillas, pero, al igual que los coeficientes de asociación al uso, cuando se quiere comparar una relación obtenida con otra distinta no hay criterios para hacerlo. Y es que «r», al contrario que la mayoría de los estadísticos, es universal.

---

[15] La proporción marginal de una modalidad se representa mediante una ecuación lineal, siendo esta proporción el resultado de la suma de las diferencias de proporciones de las modalidades independientes, respecto a una modalidad base, ponderadas por su peso más una constante.

LA BÚSQUEDA DE UNA ESTRUCTURA TOPOLÓGICA: BENZECRI

A principios de los setenta, Benzecri comienza a trabajar en una teoría más general de los datos. El análisis de correspondencias suele considerarse como una mera extensión de los trabajos de análisis factorial al caso de variables nominales. Los trabajos de Benzecri son anteriores incluso a los trabajos sobre los modelos logarítmico-lineales; sin embargo, no se traduce al inglés hasta mediados de los noventa y sus textos son desconocidos fuera del ámbito francés. Hasta entonces el conocimiento de su obra era parcial y se debía casi en exclusiva a un discípulo suyo de la Universidad de Johanesburgo, Greenacre. Entre las muchas originalidades del trabajo de Benzecri, la que más importa aquí es la consideración del Ji-cuadrado como distancia. De hecho se le suele llamar distancia o incluso métrica de Benzecri. Así, de una tabla de contingencia puede derivarse una matriz de distancias entre las combinaciones de las modalidades. Mientras que el análisis factorial clásico parte de la matriz de correlaciones, el análisis de correspondencias parte de la matriz de distancias, entre los perfiles fila y columna.

Benzecri y sus innumerables discípulos —Escofier, Pagès, Rouanet, Leroux, Lebart, Fenelon...— consiguen unificar la teoría del análisis multivariante con independencia de la medida de los datos. Benzecri lo que hace es dotar al producto cartesiano de una estructura topológica. Su planteamiento es muy inteligente: los marginales de los que Pearson obtenía las frecuencias esperadas, son convertidos en vectores —vectores fila y columna— mediante la definición de una distancia, distancia que se origina a través del «peso» de las frecuencias. Es decir, la tabla de contingencia es ahora un «espacio métrico», un conjunto dotado de una estructura, con operaciones internas y externas definidas[16].

Los resultados en el análisis de correspondencias hacen desaparecer a las variables. Los factores resultantes combinan grupos canónicos entre sí. Las nuevas variables resultantes son agrupaciones de modalidades.

---

[16] Los trabajos de Benzecri muestran una gran analogía con la mecánica newtoniana. De hecho, el sumatorio de los pesos por las distancias es nulo para que no exista movimiento. El concepto de inercia como varianza también es fundamental en el análisis de correspondencias.

Benzecri, mediante la introducción de distancia, convierte el producto cartesiano en una matriz de distancias:

$$A*B:=d(a_i b_j)$$

A partir de aquí se utilizan las técnicas de reducción de datos propias del análisis factorial mediante la extracción de valores y vectores propios.

Del análisis de correspondencias puede derivarse un nuevo planteamiento de la asociación en el sentido de interdependencia. En el análisis de correspondencias, la inercia, equivalente conceptual de la varianza, de las nubes de puntos fila o columna es igual al valor Ji-cuadrado, que es, en definitiva, la primera medida de asociación planteada. Se puede demostrar que la inercia proyectada sobre un eje vale uno cuando las columnas se asocian perfectamente con las filas. Éste es el máximo, siendo el mínimo el caso de independencia absoluta.

Sin embargo, cuando se está ante tablas muy próximas a la asociación perfecta, aparece el denominado efecto Guttman, bautizado en inglés con el término «Horseshoe», algo así como las pisadas de los cascos del caballo. Este efecto aclara el porqué los coeficientes basados en el criterio de Reducción Proporcional del Error no reconocen la asociación perfecta, y ofrecen un comportamiento no lineal o extraño en su aproximación al límite superior. En el caso de una tabla diagonal, teóricamente sólo debería existir un valor propio cuyo valor es la unidad. En la práctica aparece más de un valor propio, y por tanto más de un vector propio, siendo el segundo eje factorial una función cuadrática del primero, y el tercero una función cúbica, y así sucesivamente. De forma resumida, cuando se está una asociación perfecta los datos se proyectan sobre un espacio vectorial armónico[17]. Su homología con los conjuntos fractales también resulta evidente.

[17] La explicación del efecto Guttman es que en el primer eje las distancias entre los extremos están distorsionadas, debido a las casillas vacías que comparten. Y esto es así, porque recuérdese que la métrica de Benzecri no es euclídea.

REFLEXIÓN SOBRE LA ASOCIACIÓN

Vistos los mayores «hitos» de los trabajos sobre la asociación, las conclusiones no pueden ser más decepcionantes. Un siglo de propuestas no ha conseguido un abordaje consistente del problema. En primer lugar, el atractivo «lenguaje de las funciones» no ha permitido considerar a la tabla de contingencia formalmente como un objeto matemático específico. En segundo lugar, se ha pasado por alto la definición del concepto de asociación. Al margen de los intentos refinados de construcción de índices sintéticos, sólo los intentos de Lazarsfeld y especialmente el realizado por Benzecri han comenzado por un «lenguaje de los datos», como marco general que pudiera dar soporte a la noción de asociación. El proyecto de Lazarsfeld se ha mostrado imposible, lo desconocido sólo puede ser observado si se supone previamente. El proyecto de Benzecri, sólidamente fundado, tropieza con su tímida difusión en la literatura anglosajona y especialmente por su ausencia en los paquetes estadísticos que marcan el estándar de la investigación como el SPSS, ausencia interesada por motivos comerciales[18]. Pero su gran escollo está en que el analista sea capaz de pensar en lógicas no lineales y en métricas no euclídeas. El soporte gráfico que permite el análisis de correspondencias permite una visualización rápida de las relaciones, pero, al igual que para la interpretación de una radiografía o de una ecografía, la destreza visual, el entrenamiento y la experiencia son fundamentales para su correcta interpretación.

El debate sobre la asociación, aunque la difusión de las técnicas multivariantes lo hayan eclipsado, no es nuevo. Hyman, en contraposición a McNemar —quien pensaba que los coeficientes de rango 0-1 eliminaban la discutible interpretación de los porcentajes y sus diferencias—, denunciará a su vez las dificultades interpretativas de estas medidas acotadas: «Los límites del coeficiente de correlación son formales por naturaleza, y lo que el analista necesita para resolver el problema de si se ha acercado o no a la *mejor explicación posi-*

[18] Los trabajos de Benzecri y su escuela francesa de análisis de datos pertenecen a la filosofía del Mayo francés. Cometieron la «torpeza» de publicar los algoritmos y en vez de registrarlos. Las principales casas comerciales no tienen interés en comercializar algoritmos que puedan ser copiados por no tener derechos de autor.

ble del fenómeno, es alguna sensación más *íntima*»[19]. Hyman hablará del patrón-norma de comparación, es decir, el valor de un coeficiente no tiene sentido por sí mismo, sólo cuando puede establecerse si es mayor o menor que otro permite ser interpretado. Ahora bien, a la hora de hacer comparaciones da lo mismo hacerlas mediante coeficientes o mediante porcentajes: la conclusión es la misma y sin duda los porcentajes contienen más información —en el sentido de que están más cerca de los datos originales— que los coeficientes.

Para concluir, no podemos pasar por alto la búsqueda de «intimidad» que le preocupa a Hyman respecto a las medidas de asociación. No se trata de saber si hay mucha o poca asociación, sino también de desestimar otras posibilidades y modelos. Al paciente lector que haya llegado hasta aquí es posible que el camino recorrido le haya recordado al castigo de Sísifo. Por mi parte tan sólo he pretendido llamar la atención sobre un tema que, por banal, los textos de estadística resuelven mediante la profusión de índices de sencillo cálculo pero de oscuras interpretaciones. Mi tímida y particular respuesta es a favor de la topología de Benzecri.

BIBLIOGRAFÍA

BENZECRI, J. P. *et al.* (1973): L'analyse des données, 2 vols., Paris, Dunod.

BLALOCK, H. M. (1986): *Estadística Social*, México, FCE (e.o., 1960).

DARLINGTON, R.: *Measures of association in crosstab tables*. Documento electrónico: *http://comp9.psych.cornell.edu/darlington/crosstab/table0.htm*

DAVIS, J. A. (1976): «Analyzing contingency tables with linear flow graphs: D systems», en Heise, D. (ed.), *Sociological Methodology*, San Francisco, Jossey Bass.

DESROSIÈRES, A. (1996): «Les Apports Mutuels de la Methodologie Statistique et de la Sociologie». Comunicación presentada en *Journées de Méthodologie Statistique*, INSEE, 11-12 de diciembre de 1996. Versión electrónica en *http://www.upmf-grenoble.fr/iadest/seminaires/desros/*

GOODMAN, L. A. y KRUSKAL, W. H. (1954): *Measures of association for cross-classifications*, Nueva York, Springer Verlag.

GREENACRE, Michael J. (1984): *Theory and applications of correspondence analysis*, Londres, Academic Press.

[19] (1984: 244). El subrayado es de HYMAN.

GUTTMAN, L. (1954): «The Principal Components of Scalable Attitudes», en Lazarsfeld, P. (ed.), *Mathematical Thinking in the Social Sciences*, Nueva York, The Free Press.

HABERMAN, S. J. (1978): *Analysis of cualitative data*, 2 vols., Nueva York, Academic Press.

HYMAN, H. (1984): *Diseño y Análisis de las Encuestas Sociales*, Buenos Aires, Amorrortu (e.o. 1955).

LAZARSFELD, P. F. y HENRY, N. W. (1977): *Análisis de la estructura latente*, Madrid, Instituto de Estudios Políticos (e.o. 1968).

MCNEMAR, Q. (1955): *Psychological Statistics*, Nueva York, John Wiley.

RUIZ-MAYA, L. (dir.) (1990): *Metodología Estadística para el análisis de datos cualitativos*, Madrid, CIS.

SOMERS, R. H. (1974): «Asociación», en *Enciclopedia Internacional de las Ciencias Sociales*, vol. 4, pp. 418-422 (e.o. 1968).

YULE, G. U. (1912): «On the Methods of Measuring Association between Two Attributes», en *Journal of the Royal Statistical Society*, n.º 75, pp. 579-652.