Relaciones entre variables

Vicente Manzano Arrondo – 2013,2014

El mundo de la moda

Una de las características que define a las personas es nuestra inquietud por entender el mundo que nos rodea. Nos suele resultar interesante y positivo comprender qué ocurre y qué lo justifica. Este interés no tiene por qué tener una forma científica ni un procedimiento sistemático de investigación que lo resuelva. A pesar de ello, nos acompaña en la cotidianidad.

Aunque hasta la fecha no he prestado atención a la moda en la forma de vestir (no es una postura ideológica, sino un desinterés espontáneo), diversas conversaciones me han generado curiosidad y llevo una temporada procurando enterarme de qué pasa con ello. Escribo esto cerca del verano, lo que explica en parte lo que narro a continuación. Observo, por ejemplo, que las chicas están llevando con mucha frecuencia pantalones llamados “de pitillo”, muy ajustados, incluso mallas que suelen ser de colores muy llamativos (rayas de contraste, imitación a piel de leopardo, etc.). Con mucha frecuencia veo también a otras que llevan unos pantalones muy cortos, lo suficiente como para mostrar la parte inferior de los glúteos. Justo antes de estas semanas pre-veraniegas, cuando todavía “hacía fresquito”, observé que los pantalones estrechos terminaban en el interior de botas altas en no menos de la mitad de las transeúntes, sin demasiadas diferencias por edad. En los chicos, profusión de pantalones vaqueros anchos, que finalizan casi invariablemente en calzado deportivo que con mucha frecuencia son de color blanco o azul... Existe en ellos menos variabilidad y también menos adaptaciones al clima. Las mujeres mayores visten de una forma que depende mucho de la zona de la ciudad por donde transiten, a grandes rasgos si es más humilde o menos, o si se encuentra en la cercanía de centros de moda. En todos los casos comienzo a observar que aumenta la frecuencia de prendas de colores muy vivos, casi brillantes, que abarcan desde camisetas a zapatos. Por cierto, he quedado fascinado con una experiencia aparentemente tonta: sentarme en un banco o en un peldaño y mirar los pies de quienes pasan por mi lado. El universo del calzado se expande. La verdad es que resulta entretenido. Se trata de observación, y no sistematizada. La primera vez que compartí estos sorprendentes hallazgos, la gente que me escuchaba me mostró que son conclusiones de principiante. Existe una riqueza de matices que yo todavía no atisbo. Soy un aficionado.

Relacionando

El punto anterior pretende mostrar que las personas tenemos interés por atender a nuestro entorno y sacar conclusiones. No podemos evitarlo. Forma parte de nuestra naturaleza. Observar el contexto permite adaptarse, reaccionar o intervenir, tres reacciones que trabajan por la supervivencia.

Quienes no se fijan en cómo se viste mayoritariamente, atienden a otras cosas, a la organización de las calles, al estilo de los edificios, a la comunicación de las parejas, a la forma de conducir, a la estética de las bicicletas, al ritmo de los autobuses, a... No solo atendemos, también sacamos conclusiones. En mi caso, por lo general, soy capaz de contarte cómo creo que se sentía cada una de las cuatro personas que estaban

conversando ante mí, pero sin haber retenido ninguna característica de lo que llevaban puesto. En el ejemplo de la moda, la observación permite identificar regularidades, estableciendo conclusiones como “ahora se está llevando esto o aquello”.

Uno de los aspectos más interesantes de nuestra naturaleza de mentes inquietas se refiere al establecimiento de relaciones entre variables. Está claro que no utilizamos este vocabulario en la vida cotidiana. No acudimos a expresiones como “He observado una relación de variables. Atiende”. Más bien alguien afirma, por ejemplo, “Hoy es lunes. Cuidado con el humor de Pedro”. ¿Es esto una relación entre variables? Lo es. Observémosla más despacio.

Según la expresión, parece ser que los lunes Pedro viene con un humor difícil de soportar. El lenguaje popular está lleno de ambigüedades que, sin embargo, sirven muy bien al objetivo de comunicarnos. La frase “Cuidado con el humor de Pedro”, con el preámbulo de que hoy es lunes, está diciendo “Pedro no tiene siempre el mismo humor. En ocasiones es desagradable, especialmente desagradable, lo suficiente como para que a uno no le apetezca sufrirlo. Es lo que suele ocurrir los lunes. Así pues, si no quieres vivir una experiencia desagradable, considera alguna estrategia para sobrellevar a Pedro hoy, que precisamente es lunes”. Más allá de nuestra habilidad para comunicar muchas cosas con una expresión relativamente breve, a los objetivos de este documento interesa destacar que esa comunicación expresa una relación entre variables, aunque la persona que ha concluido tal cosa no piense ni se exprese en tales términos. Pero este es un documento de diseño y análisis de datos. Entre otras cosas, significa que hemos de practicar un estilo sistemático. Así que veamos qué significa “Hoy es lunes. Cuidado con el humor de Pedro” en términos metodológicos:

- Tenemos la variable *día de la semana*, indicada a partir de la aseveración “hoy es lunes”. Hoy es lunes, pero mañana será martes y tenemos todavía cinco valores más para la variable. Aunque sabemos que la semana tiene siete días, esta variable cuenta en la práctica con dos únicos valores: lunes y no-lunes, puesto que no hay nada en la expresión que requiera distinguir entre los seis días contenidos en la categoría no-lunes.
- A su vez, algo pasa los lunes que no ocurre el resto de los días. Es decir, el día de la semana está *co-variando* con otra cosa.
- Esa otra cosa es otra variable (si no variara, si fuera una constante, ya no tendríamos co-*variación*): *humor de Pedro*. Como además de gente de ciencia somos gente, controlamos el lenguaje humano. Esa frase está queriendo decir más de lo que captaría un programa de ordenador. Sin expresarlo literalmente, indica que Pedro está de mal humor. ¿Por qué? No lo sabemos, pero algo tiene que ver con que hoy sea lunes. Y no se diría si el humor de Pedro fuera una constante. Está claro que el resto de la semana esperaríamos un humor más agradable, un valor distinto en esa variable.
- La relación, ya lo estoy diciendo, existe porque cuando varía *día de la semana* también varía *humor de Pedro*.

Con mayor corrección científica podemos rehacer la expresión del siguiente modo: “Ser o no ser lunes y el tipo de humor de Pedro son dos variables relacionadas entre sí”. No solemos comportarnos de ese modo en la vida cotidiana. Quienes lo hacen (no conozco a nadie) podrían recibir la valoración de gente rara o pedante. Sin embargo, en el mundo de la ciencia, es así como hay que expresarse porque de otro modo nuestra comunicación sería difícil y ambigua.

Doy clases a los grupos grandes C y D. A primera hora de la mañana tenemos sesión de “grupo pequeño”, grupos formados por la cuarta parte de un grupo grande. He

observado que los grupos pequeños del C suelen venir muy poco o incluso no asisten, mientras que los grupos pequeños del D, sí, aunque tampoco en masa. Hay una posible conclusión común a todos: hay poco éxito de convocatoria. Cuando he hablado de esto con algún estudiante, me dice algo así como “Es que tus clases de grupo pequeño no son obligatorias, no puntúa la asistencia, y también nos dedicamos a cosas que no se preguntan directamente en el examen”. Fijaos que ya hemos acumulado tres relaciones en lo que va de párrafo. De forma esquemática:

Relación 1:
- Variable A: asistir más o menos a las clases de grupo pequeño.
- Variable B: pertenecer al grupo grande C o D.
- Sentido: Si varía el grupo de C a D, varía también la asistencia a grupo pequeño, en el sentido de que aumenta.

Relación 2:
- Variable A: asistir más o menos a las clases de grupo pequeño.
- Variable C: puntuar o no la asistencia.
- Sentido: si se pasa de no puntuar a sí puntuar la asistencia, esta aumenta.

Relación 3:
- Variable A: asistir más o menos a las clases de grupo pequeño.
- Variable D: abordar o no en grupo pequeño conocimientos que serán directamente evaluados en el examen final.
- Sentido: si se pasa de no a sí abordar conocimientos directamente evaluables, aumenta la asistencia.

Insisto en que podemos observar variación en dos variables, pero no encontrar *covariación*, es decir, no observar ningún indicio de que varían de forma conjunta (en alguna medida). Por ejemplo, una mosca a mi alrededor vuela con apariencia desordenada. Su posición varía, por lo que tenemos una variable: posición de la mosca en el espacio. Por otro lado, la sangre de mi cuerpo circula a una velocidad que varía según los latidos. Tenemos entonces otra variable. Las dos (posición de la mosca, velocidad de mi sangre) son variables, pero no observo ninguna relación, es decir, no hay datos, información, indicios o sospechas de que *posición de mosca* y *velocidad de sangre* varíen de forma conjunta en algún sentido.

Pasos para estudiar una relación

Interesa contar con una visión de conjunto antes de seguir. Ya sabemos qué es una relación. Veamos qué hacemos con ella. Para sacar rendimiento a un análisis de una relación entre dos variables, es muy aconsejable tener presente un proceso secuencial. En muchas ocasiones, las personas que investigan terminan cuantificando relaciones directamente sin atravesar los pasos previos, o incluso directamente se implican en procesos de inferencia. Estos comportamientos son muy poco aconsejables porque perdemos información valiosa por el camino, no estamos atendiendo a las particularidades de los datos, estamos tal vez mezclando cosas diferentes o concluyendo con menos posibilidades de generalizar los resultados. Para observar los pasos en el estudio de una relación podemos pensar en las siguientes habilidades, es decir, ser capaces de:

1. Conocer a cada variable por separado.
   Antes de ver si A y B se relacionan entre sí, debería conocerlas. Conocerlas implica atravesar los pasos que ya conocemos para los estudios unitarios: tabular o representar gráficamente la variable, depurar lo que se requiera, calcular índices específicos (como tendencia central y dispersión) e interpretar todo ello, además de identificar casos raros que exigen un tratamiento especial. Solo cuando esta etapa está superada tiene sentido sumergirse específicamente en la relación.

2. Identificar una relación.
   Es tanto como encontrar donde están esas *no menos de dos variables* (sin confundirlas con constantes que posiblemente también se encuentren en el texto) y qué elementos están indicando que ambas varían de forma conjunta. Esta habilidad participa también de otras: ser capaces de redactar una relación, sin necesidad de acudir al verbo “relacionar”. Además de otros verbos directamente asociados (influir, estar en función de, afectar a, ser dependiente de...), se puede articular redacciones sin generar redacciones tan artificiales como “las variables sexo biológico y número de cigarrillos fumados en un día están relacionadas entre sí”. Por ejemplo: “mujeres y hombres fuman cantidades diferentes de tabaco”.

3. Categorizar la relación.
   Una vez que la hemos identificado, la siguiente habilidad es saber de qué tipo es. Por ejemplo: relación de una variable cuantitativa con otra cualitativa con dos valores o categorías.

4. Describir.
   Las técnicas, ya lo sabemos, suelen agruparse en dos tipos: las que nos ayudan a conocer lo que ocurre mediante recursos de tabulación y representación gráfica, y las que cuantifican un comportamiento o faceta específica de los datos. Lo primero que hacemos con una relación es tabular o representar gráficamente. Ello nos permite describir el conjunto de los datos asociados de ambas variables. Gracias a ello, podemos establecer ya conclusiones, observar por dónde va la cosa, encontrar casos raros o particularidades que requieren un tratamiento específico, etc. Es posible que la descripción desaconseje continuar con el siguiente paso porque ya está claro lo que ocurre.

5. Cuantificar.
   Dentro de cada categoría de relación y en función de varios criterios (como es el objetivo de investigación implicado), suelen existir varias posibilidades para cuantificar la relación. En esta asignatura procuraremos abordar solo una por cada categoría de relación. El procedimiento generará un número como resultado final. El número no solo es limitado (deja mucho fuera) sino que exige saber interpretarlo. Y eso nos consumirá concentración. Hemos de ser capaces de saber qué está diciendo ese número. Para ello, acotaremos el resultado, es decir, traduciremos el número a un intervalo con un mínimo y un máximo. De este modo, podremos interpretar qué quiere decir el resultado numérico comparándolo con ambos extremos.

6. Inferir.
   Cuando la cuantía de la relación es suficiente, estemos trabajando con una muestra y no con la población, y el muestreo haya sido aleatorio, entonces

podremos poner en marcha un proceso de inferencia estadística. Este proceso permitirá concluir si existe relación en la población de la que proviene la muestra.

Cuantificación

En una situación cotidiana, la relación entre variables es sentenciada en términos cualitativos: “Esa curva genera muchos accidentes” (variable *qué curva* relacionada relacionada con la variable *número de accidentes*), “Lo mejor es comer temprano” (variable *efectos de la ingestión* relacionada con la variable *momento de la ingestión*), “Clara resuelve los problemas de cálculo mejor que nadie” (variable *quién* relacionada con la variable *calidad del resultado*). Pero en ciencia el procedimiento preferido (que no el único) es la cuantificación, es decir, expresar de forma cuantitativa el grado en que dos variables están relacionadas entre sí. Dado que estamos abordando aspectos de análisis de datos, nos centramos en esta perspectiva.

Cuantificar significa asignar un número o cuantía a algo. Si vamos a cuantificar una relación, entonces asumimos que se trata de una cuestión de grado (más o menos relación) y que ese grado puede ser expresado de forma numérica, respetando las cuantías: una relación expresada mediante el número 7 es de mayor grado que otra con cuantía 5.

Para cuantificar una relación necesitamos un procedimiento. Los procedimientos tienen nombre. Imagina que se nos ocurre un procedimiento de cuantificación de una relación al que llamamos M de Martínez. Aplicamos la M de Martínez a la relación entre la posición de la mosca y los latidos del corazón y obtenemos M = 0,1. Esto, así expresado, es como no decir nada. No podemos interpretarlo porque no sabemos cómo funciona la M de Martínez y, por tanto, qué hacer con un M = 0,1. Si te digo que llevo 250 euros en el bolsillo, puedes dejar libre tu imaginación y pensar qué podrías hacer con ese dinero. Pero si te digo que llevo 250 tunaidíes, no podrás concluir nada porque no tienes ni idea de qué moneda es esa. Tal vez 1 euro = 2500 tunaidíes, por lo que si se me pierden no parece perderse nada. Pero quizá 1 tunaidí = 314.000 euros. En tal caso, a lo mejor compartas conmigo tu sueño de viajar a Barbados o de montar la fábrica de tejidos de algodón ecológico.

Así que todos los procedimientos han de ir acompañados de criterios claros para saber qué hacer con las cuantías. Por ejemplo, nuestra M de Martínez puede funcionar así: va de 0 a 10; conforme más cerca esté de 0, menos relación; conforme más cerca esté de 10, más relación. Gracias. Con esto ya tenemos mucho, puesto que M = 0,1 es casi M = 0 y, por tanto, nos está indicando que la relación entre ambas variables es prácticamente nada. Hay que tener en cuenta que una relación de cuantía exactamente 0 es en la práctica imposible. Tal vez sea 0,03 y la hemos redondeado a 0,0. En sentido estricto, el 0 no existe en la naturaleza de las covariaciones. Siempre hay cierto ruido de fondo, cierta covariación sin ningún tipo de identidad. Esto hay que asumirlo. Por eso pedimos a las cuantías de las relaciones un mínimo para empezar a sospechar que está ocurriendo realmente algo.

En muchas ocasiones, el estudio de una relación no es el objetivo final, sino uno intermedio, un instrumento para llegar a lo que realmente interesa en la investigación. En tales casos, se precisa concluir si existe o no relación, es decir, pasar del grado o continuo a una dicotomía, antes de seguir. Para ello se requiere un punto de corte, un valor concreto que indique a partír de qué cuantía vamos a considerar la existencia de relación. Por ejemplo, si indicamos ese punto de corte en M=5, estamos indicando que no hay relación cuando $M < 5$ y sí la hay cuando $M \geq 5$.

También, en muchas ocasiones el objetivo es precisamente el estudio de esa relación pero no hay hábito en concluir “la relación es de grado o cuantía X”, sino igualmente terminar el proceso indicando si existe o no relación. En este segundo tipo de situaciones (cuando la conclusión es finalista y no intermedia o instrumental), esta decisión puede ser matizada. Es decir, podemos considerar una categorización menos simple del continuo, contemplando por ejemplo cuatro posibilidades: relación nula, baja, media y alta. Es un hábito creciente.

La cuantía de una relación recibe en ocasiones la denominación *tamaño de efecto*. Aunque a veces se guarda esa expresión para cuantías estandarizadas (medidas específicas que no vamos a abordar en este documento), aquí vamos a considerar que un tamaño de efecto es una medida de relación que está limitada o acotada, es decir, que transforma la cuantía original para expresarla dentro de un intervalo que facilita la interpretación. Por ejemplo, para estudiar la relación entre dos variables nominales se suele acudir a un proceso de cuantificación que se denomina *Chi cuadrado de Pearson*. La Chi cuadrado suministra un número que no puede ser inferior a 0, pero que no tiene límite superior. Por ello, es difícil interpretar por ejemplo la cuantía 17. No sabemos qué significa una chi cuadrado de valor 17. Para solucionarlo, contamos con la V de Cramer, un recurso que transforma la chi cuadrado en una escala que va de 0 a 1, donde 0 significa que las dos variables nominales son completamente independientes entre sí (no hay co-variación), y 1 expresa una relación máxima (cualquiera de ellas puede ser comprendida, explicada o pronosticada conociendo a la otra). Así que cada vez que vayamos a estudiar la relación entre dos variables nominales y utilicemos la Chi cuadrado de Pearson, la transformaremos con la V de Cramer para interpretar el resultado. Es decir, utilizaremos la V de Cramer como medida del tamaño de efecto.

Lo habitual será recurrir a estrategias que permitan conseguir una medida que vaya de 0 a 1, con la misma interpretación que en el caso de la V de Cramer. ¿Qué hacemos seguidamente? ¿Cómo interpretamos algo que va de 0 a 1? La respuesta es tan deseada como insatisfactoria. El sentido común aconseja no cegarse en puntos de corte. Cada situación es distinta y requiere sumergirse en ello para tomar buenas decisiones. No obstante, los “depende” son desagradables 1) cuando se trabaja con números, 2) cuando un estudiante se enfrenta a un examen, 3) cuando se tiene prisa y 4) cuando no tenemos ganas de pensar demasiado. Para salir del paso, vamos a considerar puntos de corte. Ello permitirá, entre otras consecuencias, afrontar un examen con menos ansiedad que manejando algún “depende”. Pero tengamos en cuenta que es una situación para salir del paso, no para tomárselo literalmente en serio. Voy a plantear dos situaciones para ejemplificar esto.

Pongamos que vas a lanzarte desde un avión en paracaídas. ¿Qué relación existe entre activar el resorte y que el paracaídas se abra? Pongamos M = 6. Estamos utilizando M = 5 como referente o umbral, por lo que consideras que *sí* hay relación y, por tanto, te tiras. O no... ¿Seguro que un umbral M = 5 es suficiente para ti en esta situación? Ten en cuenta que un valor M = 6 indica que no es raro que al activar el resorte, el paracaídas no se abra, con una consecuencia poco agradable. Aunque supera el punto de corte, para ti posiblemente no sea suficiente y decidas manejar, digamos, M = 9,9. En tal caso, dado que M = 6 < 9,9 entonces decides no lanzarte del avión con ese paracaídas.

Pero imagina que el avión está en llamas. Vas a morir con absoluta seguridad (si es que existe tal cosa). Sin embargo, sabes que la relación entre saltar del avión con un paracaídas y que este se abra está cuantificada en M = 4. Si seguimos en la situación inicial de M = 5, te abrasarías. Llevarías a cabo esta comparación: “como M = 4 < 5, entonces no me tiro”. Posiblemente coincidas conmigo en que sería una mala decisión. M = 4 no es M = 0 y, por tanto, no sería raro que el paracaídas se abriera y te salvaras. Pero

si te quedas, no lo cuentas. En la práctica, te serviría cualquier cosa superior a M=0. Tú te tiras seguro. Mejor M=4 que nada.

Son dos ejemplos algo forzados. Pero sirven a su cometido: mostrar que una decisión artificial y arbitraria no puede tomarse como una verdad indiscutible. Cada situación es nueva en sí misma y requiere reflexión. Prescindir de ella para dejar que un procedimiento automático tome las decisiones es algo cuando menos poco inteligente. No sé si te pasa lo mismo, pero cada vez que me dicen que entro en un edificio inteligente, o subo por un ascensor inteligente, o que mis datos están siendo tratados por un sistema inteligente... se me ponen los pelos de punta.

No obstante, repito que para salir del paso y establecer algún criterio en el que agarrarnos en situaciones donde es muy difícil identificar consecuencias, para procesos intermedios y para vencer la ansiedad a un examen, vamos a considerar tres puntos de corte como criterio automático. Para ello nos basamos en la propuesta de Cohen[1] respecto al coeficiente de correlación de Pearson (*r*, que veremos en otro documento). Cohen propuso que $r = 0{,}1$ o similar representa un efecto pequeño; en torno a 0,3 es mediano; y en torno a 0,5 es grande. Personalmente pienso que son cotas muy generosas. Como veremos, la *r* es un índice muy sensible. Podemos ver un diagrama de dispersión que no dice nada y obtener en cambio un valor $r = 0{,}3$. Pero dado que según ha quedado claro las acotaciones buscan reducir la ansiedad y orientar, pero no sustituir el sentido común ni el conocimiento de experto sobre la situación, y dado también que esas acotaciones son las más aceptadas, utilizadas y extendidas, pues no vamos a ponernos muy trascendentes con la cuestión y tomamos a Cohen como inspiración no literal. Va a ser no literal porque en lugar de “en torno a”, vamos a considerar esos valores como cotas máximas y afirmar que si B es el valor del índice concreto (en el último ejemplo, B = *r*), entonces las cotas serán:

Efecto nulo	$B \leq 0{,}1$
Efecto pequeño	$0{,}1 < B \leq 0{,}3$
Efecto mediano	$0{,}3 < B \leq 0{,}5$
Efecto grande	$B > 0{,}5$

En documentos específicos iremos abordando diferentes estrategias para cuantificar relaciones, según los objetivos y la escala de medida de las variables. Una vez entremos en contacto con un procedimiento de cuantificación de la relación, el interés se centrará en cómo transformarlo en una escala acotada en (0,1) para aplicar los criterios de la tabla anterior y poder concluir en términos de tamaño de efecto nulo, pequeño, mediano o grande.

1 Cohen, J. (1988). Statistical power analysis for the behavioral sciences. New Jersey: Lawrence Erlbaum.