El concepto de asociación estadística

Tema 6

Estadística aplicada

Por Tevni Grajales G.

En gran medida la investigación científica asume como una de sus primera tareas, identificar las cosas (características o factores) que participan en un fenómeno. Esta participación implica que existe una especie de asociación o relación entre los elementos que conforman el fenómeno.

Por ejemplo, si observamos y estudiamos el crecimiento de una planta, nos damos cuenta que hay diversos asuntos (factores, características, condiciones) que están involucrados en ese fenómeno: el tipo de planta, las características del suelo, la cantidad de agua y luz disponibles, etc. Desde el punto de vista de la estadística, se supone que podemos observar y medir estas características.

Una vez hemos medido el comportamiento de cada característica, podemos intentar identificar la forma como interactúan los cambios que sufre cada una de ellas. Continuando con el ejemplo de la planta, cuando se incrementa la cantidad de agua, ¿qué sucede con las otras características observadas? Si la respuesta es: no sucede nada, no hay cambio, es posible que estamos ante una falta de asociación o relación entre la característica observada (cantidad de agua) y las otras características. Pero cuando existe una especie de correspondencia en la forma como se manifiesta la medición de cada característica, a medida que una de ellas varía, podemos atrevernos a suponer que existe relación entre las variables.

Es muy importante recalcar que hasta esta etapa de nuestro comentario no hemos dicho qué cosa causa cuál otra. Sólo estamos señalando que un cambio en los valores de una variable parece coincidir con cambios en los valores de las otras. Así es como llegamos a conocer cuáles características están asociadas o relacionadas en el fenómeno conocido como germinación y crecimiento de una planta.

De manera que para que exista una asociación, necesitamos por los menos de dos mediciones. De la misma manera que para que haya un conflicto se necesitan, al menos, dos que estén dispuestos a pelear.

La asociación se manifiesta de diversas maneras

Hemos visto que es necesario medir el comportamiento de ciertas características para así poder determinar, con técnicas estadísticas, el grado de asociación que existe entre ellas. En algunos casos la asociación puede ser muy limitada o débil; en otros casos se trata de asociaciones fuertes.

El comportamiento de una asociación puede manifestarse de diferentes maneras: 1. Cuándo los valores de una característica aumentan, también aumentan los valores de la otra. Puede tratarse de un incremento aritmético - lineal o geométrico -curvilinear. 2. Cuando los valores de una variable aumentan los de la otra disminuyen. 3. No hay un comportamiento armónico en la forma como los valores de una y otra variable cambian (supone que no existe asociación entre las características observadas).

Más adelante entraremos en detalles respecto a la forma de observar estos comportamientos.

Los niveles de medición y las medidas de asociación

Antes de seguir adelante, es importante recordar algo que ya hemos estudiado. Recuerde que cuando

hacemos una medición, lo hacemos según el tipo de variable que se observa y según el nivel de medición que mejor corresponda. Así tenemos las variables cualitativas que conduce a la utilización de escalas nominales y/o tal vez cualidades presentadas según una jerarquía u orden y en ese caso estamos utilizando una escala ordinal. Otras variables pueden ser cuantitativas discretas o continuas las que dan lugar a la utilización de escalas de intervalos exactos y escalas de razón. Como ustedes recuerdan, los niveles de medición tienen características que les permiten o no acceder a ciertas operaciones aritméticas de manera que al tratar de determinar estadísticamente, la relación entre variables, es importante considerar con qué escala de medición se ha medido cada variable. Eso determina el tipo de técnica estadística a ser utilizada.

El hecho que los valores de las características se correspondan al observar sus variaciones no garantiza que exista una asociación lógica.

Es muy importante aclarar que al hacer comparaciones aritméticas entre dos conjuntos de datos correspondientes a sendas variables, es responsabilidad del investigador asegurarse de que existe una razón lógica que dé sentido a la asociación. Eso me recuerda lo que en cierta ocasión leí respecto a un pequeño pueblo en Europa donde la temporada del año en la que llegan las aves migratorias coincide con el incremento de los nacimientos de niños. Y es probable que al comparar el número de aves que anidan en el parque cada día del año con el número de niños que nacen en el hospital el comportamiento de las cifras manifiesten cierta correspondencia, pero eso no es suficiente para llevarnos a la conclusión de que la llegada de las aves al pueblo está relacionada con la llegada de los niños. Casi sería como decir que, en efecto, a los niños los traen las cigüeñas.

Diferentes formas de determinar la asociación.

Al tomar en cuenta el hecho de que las variables pueden ser medidas según diversas escalas o niveles de medición, inmediatamente entendemos que no se puede utilizar un procedimiento único para valorar el grado y forma de asociación entre las variables.

Cuando tenemos una variable medida a nivel nominal y esta asume más de dos valores, tenemos en mano una medida muy limitada con la cual algunos preferimos decir que no es posible determinar una relación. Por lo que se dice que lo que hacemos al estudiar la asociación con variables medidas en escala nominal, es determinar si existe dependencia entre las variables.

Esto conduce a afirmar que la asociación entre variables puede ser: 1) de dependencia, cuando se trata de probar asociación con niveles de medición nominal y 2) de relación cuando se miden las variables a nivel ordinal, intervalar o de razón. Cuando disponemos de este último tipo de mediciones, lo que hacemos es determinar lo que se denomina un coeficiente de correlación entre las variables.

Existen muchos tipos de coeficientes de correlación. La decisión de cuál es ha de emplear para un conjunto específico de datos depende de factores tales como: 1) tipo de la escala de medida en que cada variable está expresada; 2) la naturaleza de la distribución (continua o discreta); y 3) la característica de la distribución de las calificaciones (lineal o no lineal).

Tabla 1

Diversos tipos de coeficientes de correlación y escalas numéricas con que son utilizados.

ESCALA	COEFICIENTE	SE USA CON

Nominal	Chi cuadrado (no se trata de un coeficiente pero conduce a una prueba de dependencia). Lambda (λ)	Variables nominales (requiere de un número adecuado de casos)
	Coeficiente phi (r_{phi})	Dos variables dicotómicas
	r biseral (r_b)	Una variable dicotómica cuya continuidad se admite; una variable que puede tomar más de dos valores
	r tretacórica (r_t)	Dos variables dicotómicas cuya continuidad se puede admitir.
Ordinal	R de Spearman (r_s)	Datos ordenados según su rango. Si una variable es propiamente ordinal y la segunda es de intervalo/razón, se las debe expresar a las dos según su rango antes de calcular la r de Spearman
	Tau de Kendall, o coeficiente de correlación por rangos (τ)	Datos ordenados según su rango
Intervalo/razón	r de Pearson	Variables continuas o discretas medidas en intervalo o razón

En esta ocasión nos detendremos a considerar las coeficientes de correlación de Pearson (variables medidas en escala de intervalo o de razón) y de Spearman (para datos ordenados según su rango). Además daremos alguna breve explicación al uso de la chi cuadrado para determinar la dependencia entre variables.

No importa cuál sea la técnica de correlación que se utilice, hay ciertas características comunes a ellas.

1. Se utilizan dos conjuntos de medidas en los mismos individuos (sucesos) o en parejas de individuos que tengan alguna forma de relación.

2. Los valores de los coeficientes de correlación varían entre +1.00 y -1.00. Ambos extremos representan relaciones perfectas entre las variables, y .00 representa ausencia de relación.

3. Una relación positiva significa que los individuos que obtienen valores altos en una variable también tienden a obtener valores altos en la otra. La aseveración contraria también es válida; es decir, los individuos que obtienen valores bajos en una variable tienden a obtener valores bajos en la otra. (En este caso se supone que la relación entre las variables es lineal)

4. Una relación negativa significa que los individuos que obtienen valores bajos en una variable tienden a obtener valores altos en la segunda variable. (En este caso se supone que la relación entre las variables es lineal).

La relación entre dos variables es representable por medio de lo que conocemos como diagramas de dispersión. Sobre dos ejes (abscisas y ordenadas) se representa cada caso observado ubicándolo en el punto donde se encuentra el valor de la variable X medida sobre un eje y su correspondiente valor para la variable Y medido sobre el otro eje.

A continuación se presentan las figuras 1 y 2 las cuales representan la relación que se observó en una muestra de 475 empleados de las universidades adventistas de Iberoamérica. En la figura 1 se tiene la correlación entre la edad del empleado y el salario que recibe. Estas variables tienen una correlación de -.05 (sig. = .311) lo cual significa que no existe relación dado que el coeficiente casi es 0. En cambio en la Figura 2 se representa la correlación observada en los mismos empleados pero entre las variables compromiso con la institución y grado de participación. En este caso existe una correlación positiva con un coeficiente de .80 (sig. = .000) lo cual significa una correlación fuerte.

Al observar comparativamente ambos diagramas, se puede notar que la Figura 2 presenta los casos agrupados de tal manera que parece un figura ovalada que se ubica de abajo hacia arriba de izquierda a derecha (es una representación típica de una correlación positiva). Si fuese la misma figura pero con la inclinación de arriba hacia debajo de izquierda a derecha sería una correlación negativa. Volviendo la atención a la Figura 1 podemos notar que los casos se dispersan por todo el recuadro, son una forma definida. Esto es típico cuando se trata de dos variables que no están relacionadas.

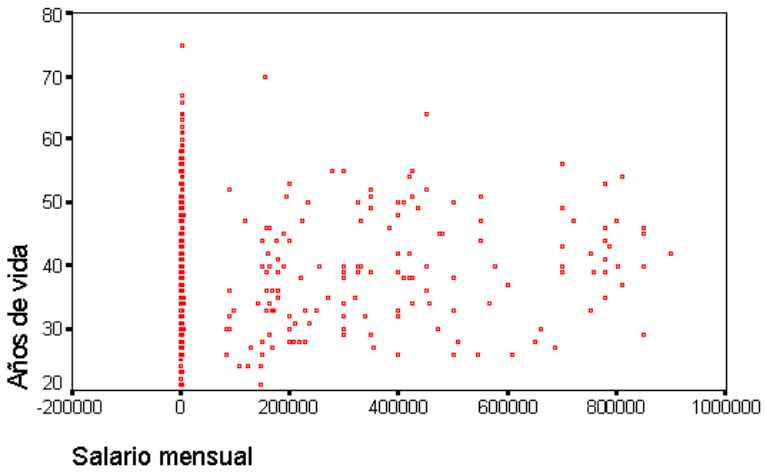

Figura 1

Diagrama de dispersión de las variables salario e edad.

Coeficiente r = -.0476

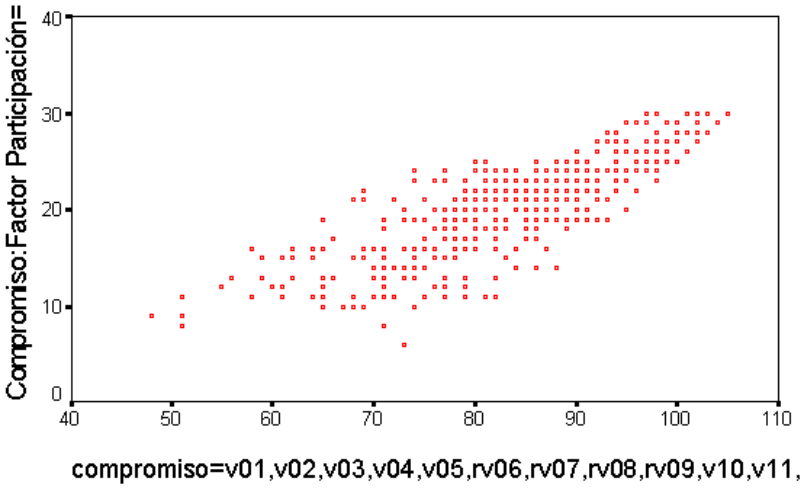

Figura 2

Diagrama de dispersión de las variables compromiso y participación.

Coeficiente r = .8046

Para explicar lo que significa una correlación tomemos el caso de la r de Pearson. Cuando tenemos un valor positivo alto de la r de Pearson como es el caso de la figura 2, esto indica que cada individuo obtiene aproximadamente, las mismas calificaciones z en ambas variables. En una correlación positiva perfecta (r = 1.00), cada individuo obtiene exactamente la misma calificación z en ambas variables. Con una r negativa alta, cada individuo obtiene aproximadamente la misma calificación z en ambas variables, pero con signos diferentes.

Entendiendo que el valor z representa una medida de posición relativa en una variable dada (es decir, un valor positivo alto de z representa una alta calificación relativa al resto de la distribución, y un valor negativo alta de z representa una baja calificación relativa al resto de la distribución) se puede generalizar el significado de la r de Pearson.

"La r de Pearson es una medida del grado en que los mismos individuos o sucesos ocupan la misma posición relativa respecto a las dos variables"(Runyon, 1992 p.126)

Tabla 2

Calificaciones originales y sus correspondientes valores z de 7 individuos con respecto a dos variables correlacionadas de manera perfecta y positiva. (datos hipotéticos)

Caso	x	x- x	$(x - x)^2$	z_x		y	y - y	$(y - y)^2$	z_y	$z_x z_y$
A	1	-6	36	-1.5		4	-9	81	-1.5	2.25
B	3	-4	16	-1.0		7	-6	36	-1.0	1.00
C	5	-2	4	-0.5		10	-3	9	-0.5	0.25
D	7	0	0	0		13	0	0	0	0

E	9	2	4	0.5		16	3	9	0.5	0.25
F	11	4	16	1.0		19	6	36	1.0	1.00
G	13	6	36	1.5		22	9	81	1.5	2.25
Σ x = 49 Media = 7.00 SC_x = 112 S_x= $\sqrt{}$ 112/7 = 4.00							Σ y = 91 Media = 13.00 SC_y = 252 S_y= $\sqrt{}$ 252/7 = 6.00		Σ ($z_x z_y$) = 7.00	

Las variables de la Tabla 2 tienen escalas de medición diferentes, una va de 4 hasta 22 y la otra va de 1 a 13 puntos. Nótese que cada caso logra obtener una valor z (Ver la columna z_x y la columna z_y) similar tanto en una medición como en la otra. Si se invirtiera el orden de las calificaciones en una de las columnas, los casos tendrían el mismo valor z pero con signo invertido lo que conduciría a una correlación perfecta pero negativa.

Una de las fórmulas para calcular el coeficiente de correlación es $r = \Sigma (z_x z_y)/N$. Como se puede observar en la medida que $\Sigma (z_x z_y)$ se acerca a O la correlación no existe (r = 0).

Para calcular la r de Pearson se cuenta con el método de la desviación de la media que es la sumatoria de la diferencia entre el valor y la media de la primera variable multiplicado por la diferencia entre el valor y le media de la segunda variable dividido entre la raíz cuadrada del producto de la suma de los cuadrados de la primera variable por la suma de los cuadrados de la segunda variable.

También existe otro método conocido como de las calificaciones originales y otra para cuando las medias y las desviaciones estándar han sido calculadas previamente. Estas fórmulas no se presentan en este documento dado que el procesador de palabras que se utiliza no ofrece la posibilidad de hacerlo de manera apropiada. (Puede conseguirla en libros es estadística recomendados en la bibliografía del curso o solicitarlo al profesor en la clase).

El coeficiente de correlación de Spearman se utiliza cuando las escalas de medición en que se han medido las variables son ordinales o de rango. Lo primero que se hace una vez ordenados los datos según sus rangos es obtener las diferencias entre los rangos que corresponde a cada caso para cada variable. Estas diferencias se elevan al cuadrado y se suman para utilizarlas en la fórmula:

$$r_s = 1 - 6\Sigma D^2 / N (n^2 - 1)$$

A continuación un ejemplo. Supongamos que siete estudiantes han terminado sus exámenes de ingreso a la universidad y se desea conocer si existe relación entre la calificación que obtuvieron en la prueba de matemáticas y la calificación obtenida en la de comunicación escrita. La información disponible en este caso no es exactamente la calificación de cada estudiante sino su ubicación con respecto a sus compañeros. La Tabla 3 presenta los datos ofrecidos por alumno y los cálculos necesarios para determinar

el coeficiente de correlación de Spearman el cual corresponde a datos ordenados.

Tabla No. 3

Orden de ubicación de 7 estudiantes en pruebas de admisión

Estudiante	Lugar en Matemática	Lugar en comunicación	Diferencia de rango (D)	Cuadrado de la diferencia (D^2)
A	1	7	-6	36
B	2	6	-4	16
C	3	5	-2	4
D	4	4	0	0
E	5	3	2	2
F	6	2	4	16
G	7	1	6	36
				$\Sigma D^2 = 110$

A fin de determinar el coeficiente de correlación de Spearman resolvemos la ecuación:

$r_s = 1 - 6\Sigma D^2 / N (n^2 - 1)$

$r_s = 1 - 6(110) / 7 (49 - 1)$

$r_s = 1 - 660 / 7 (48)$

$r_s = 1 - 660 / 336$

$r_s = 1 - 1.96$

rs = -.96

El resultado obtenido nos lleva a la conclusión que sí existe una relación entre las calificaciones obtenidas por los estudiantes, y que dicha relación es inversa o negativa, que los que obtuvieron mejores notas en matemática tienden a tener nos bajas en comunicación.

Las pruebas de dependencia

Como fue mencionado al inicio del tema, existen variables que son medidas a nivel nominal y al procurar determinar su asociación con otras variables encontramos limitaciones dado que su nivel de medición no permite utilizar con libertad todas las operaciones aritméticas. Este la forma de probar la asociación por

medio de la prueba de Ji Cuadrada será objeto de un tema posterior.

Facultad de Educación, Posgrado en Educación
Universidad de Montemorelos

Setiembre de 1999

tevgra@umontemorelos.edu.mx