LA PARADOJA DE SIMPSON Y LA INTERPRETACIÓN DE LOS RESULTADOS DE LAS EVALUACIONES DEL RENDIMIENTO ACADÉMICO EN EL SISTEMA EDUCATIVO

JOSÉ LUIS GAVIRIA (*)

Este artículo pretende alcanzar los siguientes objetivos: explicar en qué consiste la paradoja de Simpon; explicar que la paradoja de Simpson puede aparecer cuando la variable dependiente de continua; resaltar que la interpretación de los resultados de una evaluación puede ser errónea si este fenómeno se da; presentar un caso en el que dicha paradoja aparece en la recien evaluación de la Secundaría del Sistema Educativo Español, y comprobar que la correcta interpretación de los resultados puede ser contraria de lo que *prima facie* pudiera parecer.

Un médico [1] estuvo participando durante un tiempo en un experimento clínico para probar la eficacia de un nuevo tratamiento para una enfermedad de una alta tasa de mortalidad. La muestra que participó en el experimento estaba compuesta por enfermos procedentes de una zona rural y por enfermos procedentes de la capital. Al cabo del experimento el médico contaba con todos los resultados. Era además muy aficionado a las apuestas, y decidió apostar con un colega, utilizando los datos del experimento. La apuesta consistía en lo siguiente: El médico elige un tratamiento, y su colega extrae un historial clínico al azar entre las personas que han recibido ese tratamiento. Si el paciente en cuestión sobrevivió a la enfermedad, entonces el médico gana. Si el paciente falleció, entonces el médico pierde la apuesta. El colega observó que su compañero siempre apostaba por el tratamiento convencional, y no por el nuevo. Sabiendo que su adversario había participado en el estudio, supuso consecuentemente que el médico conocía los datos, y por tanto jugaba con ventaja.

Al cabo del tiempo, el médico cayó enfermo de la misma enfermedad. El colega fue a visitarle, y quedó muy sorprendido al enterarse de que había elegido voluntariamente recibir el tratamiento nuevo, en lugar del tratamiento convencional. Inmediatamente pensó que la enfermedad había afectado a su capacidad de raciocinio. Con mucha delicadeza le preguntó: «¿Cómo es que has elegido el nuevo tratamiento para tu enfermedad? ¿No habías participado en el estudio clínico correspondiente?» «Efectivamente, y es esa la razón de mi decisión. Definitivamente es el mejor tratamiento de los dos». «Pero, ¿cómo es que cuando apostábamos siempre elegías el tratamiento convencional?» «Pues

(*) Universidad Complutense de Madrid.

(1) Esta historieta es una adaptación de la presentada en BLYTH, 1972.

porque eligiendo ese tratamiento tenía más posibilidades de ganar». «Pero, ¿no quiere decir eso que ahora tienes menos probabilidades de sobrevivir?» —¡No, de ninguna manera!— Esta respuesta casi convenció al colega de que el médico tenía afectado el cerebro. Decidió no parecer grosero, y por entretener un rato al enfermo le concedió el beneficio de la duda. —¿Puedes explicarme como es posible eso?». —¡Naturalmente!», respondió el enfermo. —Basta con que observes con detenimiento la siguiente tabla, que casualmente tengo aquí (tabla I). Puedes observar que con el tratamiento convencional aproximadamente el 46% de los pacientes sobrevivieron, mientras que de los que recibieron el tratamiento experimental, sólo el 11% sobrevivieron. ¡Me pareció la opción más lógica apostar por el tratamiento convencional!». «Claro, eso lo entiendo perfectamente», –dijo el colega algo mosqueado. «De hecho sospechaba que tú tenías estos datos cuando hicimos nuestra apuesta. Sin embargo no entiendo porqué ahora que tienes que elegir tratamiento ¡no eliges también el convencional!». Lo entenderás perfectamente cuando veas la tabla II.

TABLA I

	Tratamiento convencional	Tratamiento experimental
Fallecidos	5.950	9.005
Supervivientes	5.050	1.095

Puedes comprobar que cuando se aplicaron los dos tratamientos en el medio rural, el tratamiento experimental consiguió una tasa de supervivencia de 0.1, contra 0.05263 del tratamiento convencional. En el medio urbano, el resultado fue de 0.95, contra 0.5 del convencional. Esto me convenció de que definitivamente el tratamiento experimental era el que me convenía. Es más, teniendo en cuenta que yo soy un paciente de medio urbano, y suponiendo que mis características son similares al resto de los sujetos de esa muestra, ¡mis probabilidades de supervivencia son del 95%!'

El colega se dio cuenta de que el médico estaba en sus cabales, y aunque sin acabar de entender del todo el asunto, se despidió cortésmente y pasó el resto del día dando vueltas a las dichosas tablas.

A esta curiosa situación, se la conoce como paradoja de Simpson. Ha sido descrita y tratada en varios lugares (Simpson, 1951; Blyth, 1972; Wainer, 1986). Tiene que ver con la interacción de segundo orden en las tablas de contingencia de 2x2x2.

Formalmente la paradoja de Simpson consiste en lo siguiente: Es posible que $P(a|b) \leq P(a|\bar{b})$ a pesar de que $P(a|bc) \geq P(a|\bar{b}c)$ y $P(a|b\bar{c}) \geq P(a|\bar{b}\bar{c})$. Como señala Blyth (1972), intuitivamente parece que tiene que ser que P (a | b) haya de ser algún valor intermedio entre P (a | bc) y P (a | bc), y que P (a | b) debiera ser algún valor intermedio entre P (a | bc) y P (a | bc) , y que P (a | b) debiera ser algún valor intermedio entre P (a | bc) y P (a | bc).

Como sabemos, $P(a|b) = P(a|bc)\ P(c|b) + P(a|b\bar{c})\ P(\bar{c}|b)$, y $P(a|\bar{b}) =$ P

TABLA II

	MEDIO RURAL		MEDIO URBANO	
	Tratamiento Convencional	Tratamiento experimental	Tratamiento Convencional	Tratamiento experimental
Fallecidos	950	9.000	5.000	5
Supervivientes	50	1.000	5.000	95

$(a|\bar{b}c)\ P(c|\bar{b}) + P(a|\bar{b}\bar{c})\ P(\bar{c}|\bar{b})$. Como podemos ver, el razonamiento intuitivo sería cierto, si los pesos $P(c|b) = P(c|\bar{b})$, y $P(\bar{c}|b) = P(\bar{c}|\bar{b})$. Cuando ese no es el caso, resulta que b y c no son independientes, y es su interacción la que da lugar a la paradoja mencionada.

Simpson describió este caso con tres variables dicotómicas. Es habitual que como en el ejemplo de Blyth, una de ellas actúa como variable dependiente en un experimento. Pero la paradoja también se produce cuando dicha variable dependiente es continua. Calcular las proporciones como estimadores de las probabilidades es lo mismo que en los datos de las tabla I y II asignar un cero a los casos de fallecimiento, y un uno a los de supervivencia. Tendríamos entonces para cada casilla las medias que aparecen en las tablas III y IV, donde podemos apreciar que se produce la paradoja. Nada nos impide utilizar una codificación distinta de la dicotómica, y obtener por tanto una variable dependiente continua y unas tablas similares a las anteriores en las que también se produce la paradoja de Simpson.

Es muy importante considerar la posibilidad de la aparición de la paradoja de Simpson al interpretar resultados de evaluación [2] del rendimiento. No hacerlo así llevaría necesariamente a la mala interpretación de los datos.

TABLA III

	Tratamiento convencional	**Tratamiento experimental**
Resultados	0,45909091	0,10841584

INTERPRETACIÓN DINÁMICA DE LA PARADOJA DE SIMPSON

Como hemos visto en el apartado anterior también puede darse la paradoja de Simpson cuando la variable dependiente es una variable continua. Las propiedades de la media nos ayudan a entender mejor la naturaleza del problema, y a darnos cuenta de que es posible que aparezca también en diseños en los que las variables independientes no son dicotómicas.

En la tabla V, tenemos los datos correspondientes a las medias de rendimiento en matemáticas en las pruebas de la evalua-

TABLA IV

	MEDIO RURAL		**MEDIO URBANO**	
	Tratamiento convencional	**Tratamiento experimental**	**Tratamiento convencional**	**Tratamiento experimental**
Resultados	0,05	0,1	0,5	0,95

(2) Un revisor anónimo mantiene que sólo puede hablarse de paradoja cuando los datos proceden de un experimento y no cuando proceden de una evaluación. Según el diccionario una «paradoja es una especie absurda o que lo parece». También es una «aserción inverosímil o falsa que se presenta con apariencias de verdadera». Analicemos los datos del ensayo ilustrativo. Cuando se realizan las apuestas, podemos no saber si los datos proceden de una evaluación o de un experimento. La paradoja se plantea cuando hacemos apuestas sobre esos datos. El resultado sería el mismo fuese cual fuese la procedencia de los datos. Si sólo tenemos en cuenta la tabla I, parece que el mejor tratamiento es el convencional. Si tenemos en cuenta la tabla II, nos damos cuenta de que es justo lo contrario. La consideración de la información de la tabla I en exclusiva nos empuja a hacer una aseveración aparentemente acertada, que al analizar la tabla II se demuestra errónea. En esto estriba la paradoja. No importa de donde procedan los datos.

ción de la secundaria, 1997, de los alumnos españoles de 16 años. Como resultado de la aplicación de la reforma del sistema educativo español, algunos centros ya habían implantado la nueva modalidad curricular, (Educación Secundaria Obligatoria, 4º de ESO), mientras que en otra parte importante de los mismos todavía seguían impartiendo el curriculum anterior, 2º curso de BUP. También en este grupo de edad existen otras dos vías curriculares, 2º de Formación Profesional de 1er grado, y 2º curso de REM (Reforma Experimental de las Enseñanzas Medias).

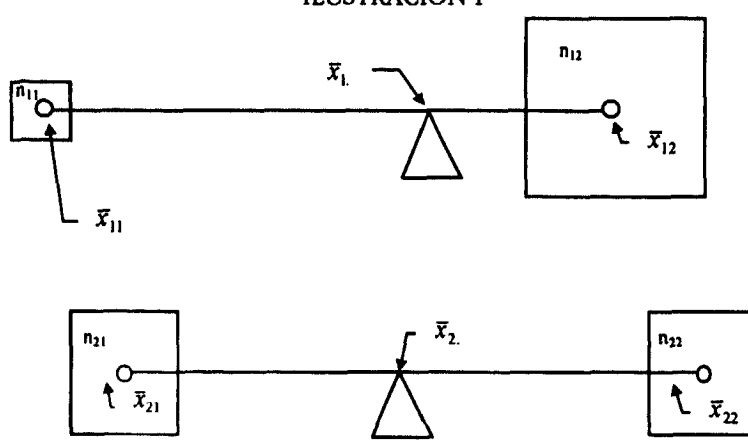
Los datos de la tabla V, corresponden a las medias de rendimiento en las pruebas de matemáticas de 4º de ESO y 2º de BUP de centros públicos y privados. Las frecuencias de cada casilla han sido variadas para ilustrar el problema. Más adelante se incluyen las frecuencias reales con las que se apreciará una variante del problema mencionado con más de dos categorías en una de las variables de clasificación.

En la ilustración I tenemos representados los datos de la tabla V. Las medias de los centros de naturaleza pública y privada se representan como puntos de una barra rígida, mientras las frecuencias representan a la masa que se apoya en esos puntos. De esta manera, si la frecuencia corresponde a la masa, entonces la media es centro de gravedad del sistema. Queda claro en esta ilustración que es perfectamente posible que las medias de 21 y 22 (4º de ESO y 2º de BUP privado) sean superiores a las de 11 y 12 (4º de ESO y 2º de BUP público) respectivamente, y que al mismo tiempo la media de la fila I sea superior a la media de la fila II.

TABLA V

	4º ESO	**Frec.**	**2 º BUP**	**Frec.**		**Media global**
PÚBLICO	268,13		278,03			276,558957
		2284		13084	15368	
PRIVADO	270,68		285,97			275,887535
		3716		1920	5636	
		6000		15004	21004	

ILUSTRACIÓN I



El que ocurra esta paradoja es posible siempre que se den ciertas condiciones. La determinación de cuáles son esas condiciones es importante en dos momentos distintos. Por una parte, cuando se está planificando un estudio; tanto de evaluación, como un experimento. En la evaluación del sistema educativo ocurre que se realizan periódicamente mediciones de los rendimientos de los alumnos con la intención de llevar a cabo comparaciones entre evaluaciones sucesivas. Cuando los instrumentos de medida son suficientemente fiables los cambios observados pueden atribuirse a cambios reales producidos en el sistema. Sin embargo los cambios en el sistema educativo se producen de forma muy lenta y paulatina. Es muy difícil que la diferencia en los valores medios de rendimiento entre dos evaluaciones sucesivas sea estadísticamente significativa. Por ello, dados los resultados de una evaluación tenemos una estimación razonable de los valores medios que se obtendrán en la siguiente en cada estrato. Sin embargo, cuando un sistema educativo se encuentra en un periodo de transición entre dos líneas curriculares distintas, en un breve periodo de tiempo nos encontramos con que la composición de los estratos de la muestra debe cambiar muy sustancialmente de una a otra evaluación. Esto nos plantea una situación en la que es posible de antemano saber si se da la posibilidad de que en la próxima evaluación aparezca la paradoja de Simpson.

También puede resultar interesante cuando se han recogido los datos y se van a interpretar. Eso nos sitúa en dos condiciones distintas. Cuando conocemos los totales marginales de una de las variables de clasificación, y cuando conocemos los cuatro totales marginales. El estudio de esas condiciones lo realizamos para el caso de dos variables de clasificación dicotómicas, pero puede generalizarse a variables con más niveles.

A. Dado el peso total de los dos conjuntos, y las medias de cada categoría en los dos conjuntos, decidir:

1. Si es posible la paradoja en esos conjuntos
2. Valores de las frecuencias en una de las categorías en uno de los conjuntos que hacen posible la paradoja.
3. Dado un valor dentro del intervalo anterior, decidir el conjunto de valores en una categoría del otro conjunto que hace posible la paradoja.

B. Dados los cuatro totales marginales, determinar:

1. Si puede haber paradoja
2. Qué valores de una casilla hacen posible la paradoja.

En los párrafos siguientes se analizan cada uno de estos puntos.

A.1 Dadas las medias de cada categoría se decide inmediatamente si existe alguna distribución de frecuencias que haga que aparezca la paradoja. Para ello tiene que darse que:

$$\bar{X}_{11} \leq \bar{X}_{21} \text{ y } \bar{X}_{12} \leq \bar{X}_{22}$$

$$min\,(\bar{X}_{21},\bar{X}_{22}) \leq max\,(\bar{X}_{11},\bar{X}_{12})$$

A.2 Dada la frecuencia total de los conjuntos, podemos fácilmente determinar en una de las categorías de uno de los conjuntos, cuáles son los valores máximos o mínimos que su frecuencia puede adoptar para que se produzca la paradoja. Supongamos que el conjunto 1 es aquél en el que las medias de las categorías son menores.

Inspeccionando la ilustración I vemos que para que la paradoja se produzca, las dos medias $\bar{X}_1$. y $\bar{X}_2$. tienen que estar comprendidas en la región determinada por $\bar{X}_{12}$ y $\bar{X}_{21}$,o en general entre $max\,(\bar{X}_{11},\bar{X}_{12})$ y $min\,(\bar{X}_{21}, \bar{X}_{22})$.

La categoría 12 puede tener el máximo de frecuencia. La categoría 11 puede tener como máximo aquella frecuencia que hace que la media $\bar{X}_1$ sea mayor que x_{21}.

$n_{11}(\bar{X}_{21}-\bar{X}_{11}) = (N_1 - n_{11})(\bar{X}_{12} - \bar{X}_{21})$ de donde

$$n_{11} \leq N_{1.}\frac{(\bar{x}_{12} - \bar{x}_{21})}{(\bar{x}_{12} - \bar{x}_{11})} \quad (1)$$

Esta es condición necesaria aunque no suficiente para que se produzca la paradoja. Si tomamos como referencia otras frecuencias, las condiciones equivalentes serían

$$n_{12} \geq N_{1.}\frac{(\bar{x}_{21} - \bar{x}_{11})}{(\bar{x}_{12} - \bar{x}_{11})}, \text{ o bien}$$

$$n_{22} \leq N_{2.}\frac{(\bar{x}_{22} - \bar{x}_{21})}{(\bar{x}_{12} - \bar{x}_{21})}, \text{ o bien}$$

$$n_{21} \geq N_{2.}\frac{(\bar{x}_{12} - \bar{x}_{22})}{(\bar{x}_{12} - \bar{x}_{21})} \quad (2)$$

A.3 Una vez determinado el valor de la frecuencia en una de las categorías de uno de los conjuntos, y puesto que hay dos totales marginales que no están determinados, entonces tenemos todavía otro grado de libertad. Supongamos que esa primera frecuencia se ha determinado en el conjunto en el que las medias son inferiores. Para que se produzca la paradoja entonces el valor mínimo que puede tomar n_{21} es tal que la medida $x_{2.}$ sea menor que $x_{1.}$.

$n_{21}(x_{1.} - x_{21}) = n_{22}(x_{22} - x_{1.})$ y $n_{22} = N_{2.} - n_{21}$

De (2) se deduce que

$$n_{21} = N_{2.}\frac{(\bar{x}_{22} - \bar{x}_{1.})}{(\bar{x}_{22} - \bar{x}_{21})} \text{ y como}$$

$$\bar{x}_{1.} = \frac{(n_{11}\bar{x}_{11} + n_{12}\bar{x}_{12})}{N_{1.}} \text{ se deduce que}$$

para que se produzca la paradoja tiene que ocurrir que

$$n_{21} \geq N_{2.}\frac{(\bar{x}_{22} - \bar{x}_{12})}{(\bar{x}_{22} - \bar{x}_{21})} - n_{11}\frac{N_{2.}(\bar{x}_{11} - \bar{x}_{12})}{N_{1.}(\bar{x}_{22} - \bar{x}_{21})} \quad (3)$$

o lo que es lo mismo, que

$$n_{22} \leq N_{2.}\frac{(\bar{x}_{12} - \bar{x}_{21})}{(\bar{x}_{22} - \bar{x}_{21})} - n_{11}\frac{N_{2.}(\bar{x}_{12} - \bar{x}_{11})}{N_{1.}(\bar{x}_{22} - \bar{x}_{21})}$$

Alternativamente, si el valor que fue determinado en primer lugar hubiese sido del conjunto 2, entonces los valores correspondientes del conjunto 1 serían

$$n_{11} \leq N_{1.}\frac{(\bar{x}_{12} - \bar{x}_{21})}{(\bar{x}_{12} - \bar{x}_{11})} + n_{22}\frac{N_{1.}(\bar{x}_{21} - \bar{x}_{22})}{N_{1.}(\bar{x}_{12} - \bar{x}_{11})}$$

o bien

$$n_{12} \geq N_{1.}\frac{(\bar{x}_{21} - \bar{x}_{11})}{(\bar{x}_{12} - \bar{x}_{11})} - n_{22}\frac{N_{1.}(\bar{x}_{21} - \bar{x}_{22})}{N_{2.}(\bar{x}_{12} - \bar{x}_{11})}$$

Naturalmente si se pretende evitar la paradoja los valores de las frecuencias deberían ser los complementarios de las cotas indicadas.

B. Un problema distinto es el que se plantea cuando los totales marginales están completamente determinados. Es el caso en el que ya se han recogido todos los datos de una evaluación, o de un experimento. Entonces podemos determinar:

1. Si puede haber paradoja con esos totales marginales.
2. Qué valores de frecuencia de una casilla, ya que sólo hay un grado de libertad, hacen posible la paradoja.

B.1 Para lo primero, suponemos que $N_1 \geq N_{2.}$. Entonces n_{11} tiene que ser suficientemente pequeño. Pero dados los totales marginales, n_{11} tiene que ser mayor o igual a $N_1 - N_{2.}$ ya que $N_1 = n_{11} + n_{21}$, y n_{21} no puede ser mayor que $N_{2.}$.

En consecuencia una condición necesaria para que se pueda producir la paradoja es que

$$(N_{.1} - N_{2.}) \leq N_{1.}\frac{(\bar{x}_{12} - \bar{x}_{21})}{(\bar{x}_{12} - \bar{x}_{11})}.$$

Y en general entonces n11 tiene como límites los dos extremos de la anterior desigualdad.

B.2 En el segundo punto, si los totales marginales permiten la paradoja, se trata de ver cuál es el valor máximo que puede tomar n_{11} para que ésta de hecho se produzca.

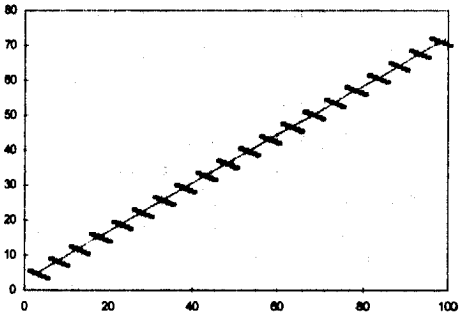
El valor más pequeño de la frecuencia de la casilla 11 está determinada por el valor más grande que puede haber en la frecuencia de 21. Y eso es el $\min(N_{.1}, N_{2.})$. Una vez determinada $\bar{x}_{1.}$ se produce la paradoja si $x_{2.} < x_{1.}$, es decir como $n_{21} = N_{.1} - n_{11}$, sustituyendo en la (3), y operando obtenemos que para que se produzca la paradoja

$$n_{11} \leq \frac{N_{.1} - N_{.2} \dfrac{(\bar{x}_{22} - \bar{x}_{12})}{(\bar{x}_{22} - \bar{x}_{21})}}{1 - \dfrac{N_{.2}}{N_{.1}} \dfrac{(\bar{x}_{11} - \bar{x}_{12})}{(\bar{x}_{22} - \bar{x}_{21})}}$$

Naturalmente para que ésta no se produzca el valor de n_{11} tiene que ser su complementario.

G.H. Haggstrom, (citado en Blyth, 1972) señala que la paradoja de Simpson es la forma más simple de la paradoja de la falsa correlación. «Un dominio x está dividido en cortos intervalos, en cada uno de los cuales y es una función lineal de x con pendiente negativa. Pero los segmentos van subiendo paulatinamente cuando avanzamos hacia la derecha, de forma que cuando consideramos la relación global entre x e y, y es prácticamente una función lineal de x con una fuerte correlación positiva.» (Véase la ilustración II).

ILUSTRACIÓN II



OTROS PROBLEMAS DE INTERPRETACIÓN EN DISEÑOS 2 x 2

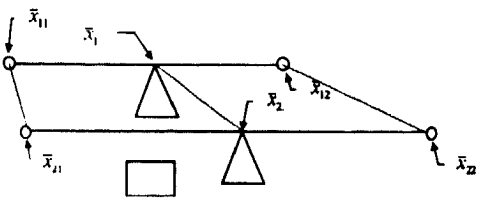
La paradoja de Simpson no es el único caso en el que puede ser un error realizar comparaciones de las medias globales. Como regla general, siempre que exista una interacción entre al menos dos variables de clasificación, la diferencia entre las medias globales de las categorías de una de las variables puede enmascarar los verdaderos resultados.

Si tenemos dos variables de clasificación, A y B, con dos niveles cada una, A_1 y A_2 y B_1 y B_2, existe interacción si la diferencia de medias en la variable dependiente entre los niveles de A_1 y A_2 es de distinta magnitud en el nivel B_1 que en el nivel B_2.

	A1	A2
B1	11	12
B2	21	22

Si se da interacción entre las variables de clasificación, ésta puede ser básicamente de dos tipos.

ILUSTRACIÓN III

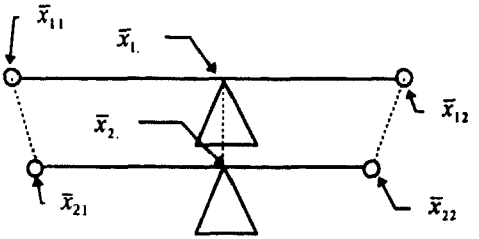


En el primer tipo las diferencias entre las categorías varían en su magnitud, pero no en el sentido. Así por ejemplo en la ilustración III, vemos que $\bar{x}_{21} > \bar{x}_{11}$, y $\bar{x}_{22} > \bar{x}_{12}$, aunque la diferencia entre las dos últimas es mayor que entre las dos primeras. En este caso la diferencia entre las medias globales, $\bar{x}_{1.} - \bar{x}_{2.}$ no representa de manera exacta la magnitud de las diferencias entre los pares de medias de las columnas, pero sí que mantiene el sentido. Es decir, al interpretar estos resultados podemos aceptar que la fila II superior a la I, aunque no podemos hacer conjeturas acerca de las magnitudes de esas diferencias, excepto en su valor medio.

En el segundo tipo de interacción las diferencias entre las medias cambian de sentido al cambiar de una a otra categoría de una de las variables de clasificación.

Podemos ver en la ilustración IV que $\bar{x}_{21} > \bar{x}_{11}$, pero $\bar{x}_{22} < \bar{x}_{12}$. En este caso la diferencia $\bar{x}_{1.} - \bar{x}_{2.}$ no representa ni la magnitud ni el sentido de los verdaderos efectos.

ILUSTRACIÓN IV

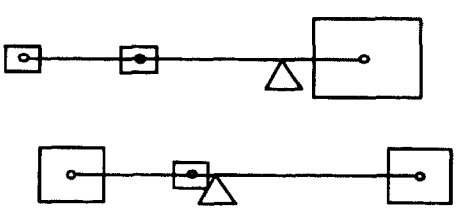


La paradoja de Simpson es un caso similar al que se da en la interacción del tipo 1, la de la ilustración III, y tiene mucha importancia porque puede darse cuando no hay interacción en absoluto. En ambos casos lo más prudente es realizar las comparaciones con los datos desagregados.

GENERALIZACIÓN A MÁS DE DOS CATEGORÍAS

A partir de la ilustración I se entiende fácilmente que la generalización a más de dos categorías en una de las variables de clasificación es inmediata. En la ilustración V vemos cómo sería un caso con tres niveles en una de las variables. En una tabla de este tipo tendríamos cuatro grados de libertad si sólo dos totales marginales estuviesen determinados, o bien dos grados de libertad si estuviesen definidos los cinco totales marginales.

ILUSTRACIÓN V

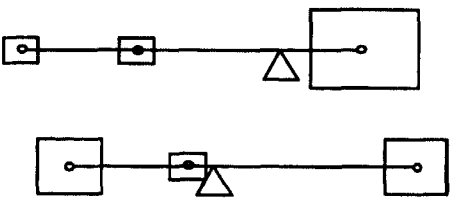


Dadas las medias de cada casilla, el valor de las cotas de las frecuencias es muy similar al caso de A.2 y A.3.

CUASI-PARADOJA

En el caso en el que una de las variables de clasificación tiene tres o más categorías, puede darse una situación como la de la ilustración VI.

ILUSTRACIÓN VI



En ella uno de los conjuntos tiene una media superior en dos categorías e inferior en una tercera. Se trata por tanto de un caso en el que puede existir interacción entre las variables de clasificación. Llamamos a esta situación cuasi-paradoja ya que no es la magnitud de las diferencias de las medias la que determina por sí sola la situación. Aparentemente esta situación no es tan llamativa como el verdadero caso de Simpson. Sin embargo, tanto en un caso como en otro la media general está enmascarando los verdaderos efectos.

Es evidente que cuando hay interacción entre las variables de clasificación, es desaconsejable la utilización de las medias generales. Pero no es el único caso. Como hemos visto con la paradoja y la cuasi-paradoja de Simpson, es posible que no haya interacción entre las variables y que sin embargo sea desaconsejable la presentación de los datos agregados.

ALGUNOS EJEMPLOS EN LA EVALUACIÓN DEL RENDIMIENTO EN SECUNDARIA EN EL SISTEMA EDUCATIVO ESPAÑOL [3]

En el momento en el que se evaluó la enseñanza secundaria en España, (1997), en los cursos modales [4] correspondientes a 16 años, coexistían cuatro modalidades curriculares con distintos niveles de implantación. Se trata de 2º de BUP, 4º de ESO, 2º de REM y 2º de FP1. Las modalidades de 2º de BUP y 2º de REM estaban en proceso de desaparición. Pero mientras la última sólo subsistía residualmente en una de las comunidades evaluadas, 2º de BUP era todavía la opción mayoritaria en la mayoría de ellas, mientras que paulatinamente se estaba implantando 4º de ESO.

Cuando se analizan los datos de los alumnos de 16 años en las distintas materias, (tabla VI), vemos que la diferencia es estadísticamente significativa a favor de los centros públicos, y sólo en gramática y literatura esa diferencia es a favor de los centros privados.

Cuando analizamos con más detalle estos datos vemos que de las cinco materias evaluadas, en cuatro de ellas, cuando hay diferencias significativas es favor de los centros privados (tablas VII a XI). En la materia restante, matemáticas, mientras que la diferencia aparente a nivel global era de 3,28 puntos a favor de los centros públicos, al desglosar los datos aparece una diferencia de 7,05 puntos a favor de los privados en segundo de BUP, de 4,34 a favor de los públicos en 2º de FP, y una diferencia de 3,64 puntos a favor de los públicos significativa sólo al nivel del 5% en 4º de ESO.

El efecto es ciertamente espectacular en ciencias de la naturaleza y en geografía e historia. En la primera de estas dos materias la diferencia aparente es de 5,68 puntos a favor de los centros públicos. Cuando analizamos por líneas curriculares vemos que las diferencias significativas son de 10,25 puntos y de 15,34 puntos ¡a favor de los centros privados!

Algo similar ocurre en geografía e historia. A una diferencia aparente no significativa de 0,33 puntos a favor de los centros públi-

TABLA VI

	Dif. Significativa A favor de:			
	Públicos	**Privados**	**Diferencia**	
Comprensión Lectora	271,46	270,36	1,1	Públicos
Gramática y Literatura	266,24	269,25	-3,01	Privados
Matemáticas	264,25	260,97	3,28	Públicos
Ciencias de la Naturaleza	268,87	263,19	5,68	Públicos
Geografía e Historia	269,9	269,57	0,33	No

(3) Ejemplos tomados de De la Orden y otros (1998).

(4) Curso modal correspondiente a una edad es aquél en el que la mayoría de los alumnos tienen esa edad.

TABLA VII

Comprensión lectora

	Públicos	Privados	Diferencias	Dif. Significativa a favor de:
2º BUP	281,68	286,34	-4,66	Privados
4º ESO	273,1	271,35	1,75	No
2º FP	246,22	246,27	-0,05	No

TABLA VIII

Gramática y literatura

	Públicos	Privados	Diferencias	Dif. Significativa a favor de:
2º BUP	289,26	298,74	-9,48	Privados
4º ESO	264,96	265,32	-0,36	No
2º FP	221,72	2226,76	-5,04	Privados

TABLA IX

Matemáticas

	Públicos	Privados	Diferencias	Dif. Significativa a favor de:
2º BUP	276,86	283,91	-7,05	Privados
4º ESO	267,27	263,63	3,64	Públicos (Sólo al5%)
2º FP	230,26	225,92	4,43	Públicos

TABLA X

Ciencias de la Naturaleza

	Públicos	Privados	Diferencias	Dif. Significativa a favor de:
2º BUP	271,21	281,46	-10,25	Privados
4º ESO	273,79	289,13	-15,34	Privados
2º FP	226,46	225,73	0,73	No

TABLA XI

Geografía e Historia

	Públicos	Privados	Diferencias	Dif. Significativa a favor de:
2º BUP	278,92	289,92	-11	Privados
4º ESO	271	281,7	-10,7	Privados
2º FP	232,3	232,91	-0,61	No

cos corresponden dos diferencias muy significativas de 11 y de 10,7 puntos en 2º de BUP y 4º de ESO respectivamente a favor de los centros privados.

En la tabla XIII aparece la distribución de frecuencias de cada estrato que dio como resultado los efectos paradójicos reseñados.

En resumen, una vez analizadas con detenimiento las líneas curriculares vemos que de diez diferencias estadísticamente significativas ocho son a favor de los centros privados, y dos, una de ellas sólo al nivel del 5%, a favor de los centros públicos.

Es evidente que las conclusiones a las que se puede llegar son radicalmente distintas cuando se tiene en cuenta la existencia de la paradoja de Simpson.

También a un nivel mayor de desagregación se producen interacciones que es necesario atender para explicar el verdadero significado de los datos. Mientras que las diferencias observadas por materia y curso entre centros públicos y privados en cada una de las comunidades autónomas es consistente [5] con los datos de las tablas VII a XI en comprensión lectora, gramática y literatura, ciencias de la naturaleza y geografía e historia, en matemáticas de 4º de ESO se da también una interacción con la variable comunidad autónoma. Así vemos en la tabla XII que hay cinco comunidades en las que la diferencia es a favor de los centros públicos, y seis en las que las diferencias son a favor de los centros privado. Además, esas diferencias varían desde 25 puntos a favor de los centros privados (Asturias), hasta 17 puntos a favor de los públicos (Cantabria) (tabla XII). Todo son claros ejemplos de cómo quedan enmascarados los verdaderos efectos si se agregan indiscriminadamente los datos.

Dado que el nivel de implantación del curso 4º de la ESO es muy distinto en unas y otras comunidades, está claro que no puede hacerse una comparación global entre centros públicos y privados referidos a este curso.

TABLA XII

Matemáticas en 4º de ESO por comunidades autónomas

GRUPO DE EDAD	COMUNIDAD AUTÓNOMA DEL CENTRO	TITULARIDAD	MATEMÁTICAS 4º DE ESO
			Media
16 AÑOS	ARAGÓN	PÚBLICO	286,94
		PRIVADO	281,71
	ASTURIAS	PÚBLICO	265,24
		PRIVADO	301,21
	BALEARES	PÚBLICO	257,86
		PRIVADO	264,81

(5) Sólo en Baleares, en Extremadura (Comprensión Lectora y Matemáticas), y en Navarra (Matemáticas), ocurre que sean los centros públicos superiores significativamente a los privados en 2º de BUP. En las dos primeras comunidades se da también el caso de que en ellas los centros privados son superiores a los públicos en todas las materias evaluadas en 4º de ESO, lo que está señalando probablemente diferencias en el ritmo de implantación de la ESO entre estas comunidades y las demás. En el caso de Navarra no se había implantado 4º de ESO en ningún centro en el momento de la evaluación.

TABLA XII (cont.)

Matemáticas en 4º de ESO por comunidades autónomas

	CANTABRIA	PÚBLICO PRIVADO	272,11 255,75
	LA MANCHA	PÚBLICO PRIVADO	270,01 282,36
	CASTIL.LEÓN	PÚBLICO PRIVADO	284,56 276,28
	CATALUÑA	PÚBLICO PRIVADO	254,79 255,18
	CEUTA/MELILLA	PÚBLICO PRIVADO	243,93 266,86
	EXTREMA	PÚBLICO PRIVADO	265,06 288,28
	GALICIA	PÚBLICO	247,32
	LA RIOJA	PÚBLICO PRIVADO	288,47 286,68
	MADRID	PÚBLICO PRIVADO	273,67 262,88
	MURCIA	PÚBLICO PRIVADO	257,59 246,02
	VALENCIA	PÚBLICO	252,10

TABLA XIII

Número de alumnos por línea curricular y titularidad del centro[6]

Curso	**Titularidad del centro**	**Alumnos en Comp. Lect., Gram. y Lit. y Mat.**	**Alumnos en CCNN y Geo. e Hist.**
2º BUP	Públicos	6852	622
	Privados	3920	479
4º ESO	Públicos	8238	1541
	Privados	927	57
2º FP	Públicos	3257	214
	Privados	2635	273

(6) Las frecuencias reseñadas son las resultantes de ponderar el tamaño del estrato por el peso que a dicho estrato le corresponde como corrección por submuestreo.

CONCLUSIONES

Cuando se presentan resultados de la evaluación del rendimiento de los alumnos en el sistema educativo, es importante poder determinar las diferencias asociadas con ciertas variables de clasificación. Por esta razón el diseñar la muestra se estratifican los datos en función de las variables consideradas de más interés. En la presentación de los resultados se tiende a hacer hincapié en los efectos principales de las variables de clasificación, y en algunas interacciones de primer orden, por razones de parsimonia. Es sabido que si existe una interacción entre las variables de clasificación, la presentación de los efectos principales en exclusiva puede enmascarar los verdaderos efectos. Pero incluso cuando no hay interacción entre las variables de clasificación respecto de la variable dependiente, pueden darse casos como el de la paradoja de Simpson en los que la interacción se da respecto de las frecuencias. Tanto en la paradoja de Simpson en diseños de 2x2, 2x3 ó más, como en lo que hemos denominado cuasi-paradoja de Simpson, en diseños de 2x3 ó más, la presentación de sólo los efectos principales enmascara el verdadero sentido de los efectos primafacie (aparentes) de las variables de clasificación.

Se han presentado las condiciones en las distribuciones de frecuencias que posibilitan la aparición de la paradoja en los diseños 2x2, y se han ilustrado los efectos de la cuasi-paradoja con datos tomados de la evaluación del rendimiento académico en el sistema educativo español.

BIBLIOGRAFÍA

BLYTH, C. R. (1972) «On Simpson's Paradox and the Sure-Thing Principle». *Journal of the American Statistical Association,* V. 67, 338, pp. 364-366

ORDEN HOZ, A. de la y otros (1998) 2. «Los resultados escolares». *Diagnóstico del sistema educativo 1997. Estudios e informes INCE,* Ministerio de Educación y Cultura. Madrid.

SIMPSON, E. H. (1951) «The interpretation of Interaction in Contingency Tables». *Journal of the Statistical Society, Series B, V.* 13, pp. 238-241

WAINER, H. (1986) «Minority Contributions to the SAT Score Turnaround: An example of Simpson's Paradox». *Journal of Educational Statistics,* V. 11, 4, pp. 239-244.