Variables y contextos en los problemas de correlación: Un estudio de libros de texto

María M. Gea[1], Carmen Batanero[2], José M. Contreras[3] & Gustavo R. Cañadas[4]

Resumen

En este trabajo se presenta un estudio empírico de los problemas de correlación presentados en una muestra de libros de texto españoles de Bachillerato. Se analizan la intensidad y sentido de la correlación, los contextos empleados y las representaciones gráficas utilizadas, deduciendo un uso sesgado de estas variables en la muestra de textos analizada. Finalmente se presentan algunas implicaciones para la investigación y la enseñanza.

Palabras clave: Correlación; tareas de estimación; variables de tarea; sesgos y concepciones erróneas

Abstract

In this work we describe an empirical study of correlation problems in a sample of high school Spanish textbooks. We analyze the correlation strength and sign; the context used in the problems and the graphical representations of correlation and deduce some biases in the use of these variables. We finally present some implications for research and teaching,

Keywords: Correlation; estimation tasks; task variables; biases and misconceptions

I. Introducción

El razonamiento sobre la correlación es una competencia esencial en la formación de los ciudadanos (Moritz, 2004; Zieffler, 2006), pues ayuda a la toma de decisiones en múltiples situaciones cotidianas y profesionales. Su importancia lleva al profesorado *"a interesarse por el razonamiento de los estudiantes en cuanto a la lectura de diagramas de dispersión, a la interpretación de la correlación y otras destrezas que son utilizadas en el estudio e interpretación de los datos bivariados"* (Zieffler, 2006, p. 7). La relevancia del tema también se justifica con las siguientes razones (Batanero, 2001):

- Permiten modelizar las relaciones entre variables estadísticas; junto con la regresión extienden la dependencia funcional, donde a cada valor de una variable

[1] Universidad de Granada, España. mmgea@ugr.es

[2] Universidad de Granada, España. batanero@ugr.es

[3] Universidad de Granada, España. jmcontreras@ugr.es

[4] Universidad de Granada, España. grcanadas@ugr.es

independiente *X* corresponde un solo valor de la variable dependiente *Y*. En la correlación y regresión, para cada valor de la variable independiente tenemos una distribución de valores de la variable dependiente. Además, se puede dar una medida de la intensidad de la relación por medio de diversos coeficientes (que dependen del tipo de variable y el tipo de relación).

- Es base de muchos otros métodos estadísticos: en análisis multivariante la correlación es una noción fundamental, pues representa el coseno del ángulo que forman dos variables, en la interpretación geométrica de los datos multivariantes. Más aún, tanto el coeficiente de correlación y la proporción de varianza explicada (cuadrado de dicho coeficiente) sirven de base a los modelos de análisis de varianza univariantes y multivariantes.

En España, este tema se estudia en el primer curso de Bachillerato (16-17 años) en las dos modalidades en que se enseña matemáticas: *Ciencias y Tecnología* y *Humanidades y Ciencias Sociales* (MEC, 2007), por lo que el tema debiera ser familiar a los alumnos universitarios. A pesar de ello, la investigación en didáctica de la matemática y psicología indica que muchos alumnos entran en la universidad sin alcanzar una comprensión profunda de este concepto, y cometen errores frecuentes en la estimación de la correlación (Sánchez Cobo, 1999).

Con objeto de contribuir a la mejora de esta situación, en este estudio analizamos las tareas que se plantean al estudiante en los libros de texto diseñados para su enseñanza en este nivel educativo, tratando de aportar directrices para su mejora ya que, de modo indirecto, sirven de apoyo para su enseñanza.

II. Antecedentes

Son muchos los autores que han destacado la importancia del libro de texto como material didáctico, y su influencia en las decisiones de los profesores sobre las tareas a realizar con los estudiantes (Stylianides, 2009). Cordero y Flores (2007) indican que el discurso matemático escolar es determinado con frecuencia por el libro de texto, que además, prácticamente regula la enseñanza y el aprendizaje.

Aunque hay una amplia investigación sobre los libros de texto de matemáticas, esta tradición es mucho menor en el caso de la estadística y probabilidad, donde encontramos algunos ejemplos, como los de Ortiz (1999), Cobo y Batanero (2004), o Azcárate y Serradó (2006).

El primer antecedente relacionado con la correlación y regresión es el de Sánchez Cobo (1999) quien analiza once libros de texto de tercer curso de Bachillerato, publicados desde 1987 hasta 1990. Como consecuencia, ofrece una taxonomía de definiciones y un análisis de las demostraciones, desde el punto de vista de la función que realizan y las componentes que la integran. Muestra una tendencia formalista en la presentación del tema, un uso mayoritario de ejemplos basados en representaciones gráficas, y un fuerte sesgo en los ejemplos presentados

hacia la correlación positiva. En su estudio, Sánchez Cobo (1999) analiza el contenido matemático, centrándose en los aspectos procedimentales (construcción de la tabla de frecuencias, cálculo distribuciones marginales o condicionales, cálculo de momentos, cálculo del coeficiente de correlación, representación gráfica y cálculo de los coeficientes de regresión), aunque no clasifica específicamente los campos de problemas, por lo que este será un punto original de nuestro estudio.

Más recientemente Lavalle, Micheli y Rubio (2006) analizan la correlación y regresión en siete libros de texto argentinos de Bachillerato, observando un enfoque mayoritariamente socio-constructivista, con un nivel de profundidad adecuado, donde se plantean más actividades bajo una asociación directa que inversa, y ninguna bajo la propuesta del uso de recursos tecnológicos.

Para complementar los citados trabajos, analizaremos las situaciones-problemas propuestas para el estudio de la correlación en textos españoles de Bachillerato. En lo que sigue se describen el método y los resultados del estudio.

III. El estudio empírico

Se analizaron ocho libros de textos de primer curso de Bachillerato en la modalidad de *Humanidades y Ciencias Sociales* (MEC, 2007), publicados recién implantado el currículo actual (MEC, 2007), que se eligieron por ser los más utilizados en la enseñanza pública, y ser de editoriales de gran tradición y prestigio en España (ver Anexo, donde se listan los libros por orden de editorial, junto con un código que se usa en el resto del trabajo). Además, esta modalidad de Bachillerato es la que da un mayor peso a la estadística en este nivel educativo.

Para cada uno de estos libros se analizan las situaciones-problema que se presentan, ya sean ejemplos, ejercicios resueltos o a resolver por el estudiante, tanto en el desarrollo del tema como al finalizar su enseñanza. Para ello se consideraron las siguientes variables, que influyen en la dificultad de una tarea de correlación (Estepa y Batanero, 1996; Sánchez Cobo, Estepa y Batanero, 2000):

- *El signo de la correlación*, que puede ser positivo en caso de dependencia directa, negativo en caso de dependencia inversa o nula en caso de independencia. Aunque matemáticamente la dependencia directa o inversa sean semejantes, los estudiantes no las perciben de igual modo, sino que tienen más facilidad en estimar correctamente las correlaciones positivas. En el trabajo de Erlick y Mills (1967) se expone que la asociación negativa se estima como muy próxima a cero.
- *Intensidad de la dependencia*, siendo más fácil por los estudiantes percibir una correlación fuerte. De hecho, los estudiantes presentan dificultades al comparar diferentes valores del coeficiente de correlación distintos de: -1, 0 y 1 (Sánchez-Cobo, 1999).

También se analizaron los contextos empleados en cada problema, y de una forma más global, la variedad de representaciones gráficas empleadas en el tema. En lo que sigue se presentan y discuten los resultados.

IV. Resultados

Intensidad

En primer lugar (Tabla 1), se clasificaron las tareas analizadas atendiendo a los distintos grados de correlación que presentan los datos: independencia, dependencia estadística alta media y baja y dependencia funcional perfecta. Así es que, cuando los datos presentan dependencia aleatoria de tipo lineal, clasificamos la intensidad de la relación atendiendo al valor absoluto del coeficiente de correlación lineal en alta, si corresponde al intervalo [0,8; 1); media, si corresponde al intervalo [0,5; 0,8); y baja, si corresponde al intervalo [0,1; 0,5). En caso de una dependencia no lineal, se considera intensidad alta si el ajuste de los datos a la función que los modeliza es muy bueno, y en caso de poder determinar el coeficiente de determinación, cuando éste sea mayor o igual a 0,98; media para valores del mismo comprendidos entre 0,8 y 0,98 y siempre que los datos se ajusten al modelo, mejorando al modelo lineal al menos en dos décimas en dicho coeficiente de determinación; y baja en otros casos.

Consideramos “independencia” para correlaciones menores a 0,1 y dependencia funcional cuando el ajuste de los datos a una función es perfecto; añadimos la categoría “descripción verbal” para aquellas tareas en que las variables se describen verbalmente o se plantean mediante expresiones algebraicas, y no es posible deducir la intensidad de la correlación.

Tabla 1. Frecuencias (y porcentajes) de tareas según intensidad de la relación

	L1	L2	L3	L4	L5	L6	L7	L8
Independencia	2(0,7)	5(2,3)	24(9,4)	12(5,3)	5(1,6)	6(3,4)	30(7,5)	21(7,1)
Baja	25(9,3)	19(8,6)	20(7,8)	52(23,1)	31(9,7)	15(8,5)	89(22,1)	27(9,2)
Media	53(19,8)	23(10,4)	34(13,3)	45(20)	55(17,3)	21(11,9)	65(16,2)	50(17)
Alta	127(47,4)	168(76)	122(47,7)	88(39,1)	169(53,1)	107(60,8)	114(28,4)	143(48,6)
Funcional	15(5,6)	4(1,8)	22(8,6)	14(6,2)	7(2,2)	3(1,7)	35(8,7)	24(8,1)
D. Verbal	46(17,2)	2(0,9)	34(13,3)	14(6,2)	51(16)	24(13,6)	69(17,2)	29(9,9)
Total	268	221	256	225	318	176	402	294

En la Tabla 1 observamos que la mayoría de las situaciones presentan una correlación alta, seguido de tareas con intensidad media, siendo menos frecuente la dependencias bajas, independencias o dependencia funcional. Los resultados coinciden con los de Sánchez Cobo (1999). En el estudio de Sánchez Cobo, Estepa y Batanero (2000) se indica que la precisión de las estimaciones del coeficiente de correlación es mayor cuando la correlación es más intensa; es posible que sea este hecho el que lleva a los autores de los textos a incidir tanto en la correlación

alta. Sin embargo, en los estudios estadísticos, especialmente en las ciencias sociales, las correlaciones suelen ser moderadas o bajas; por lo que sería conveniente una mejor distribución de la intensidad de la correlación en las situaciones-problema presentadas en los textos.

Sentido

Se ha estudiado también el sentido de la dependencia, que se clasifica en directa si el crecimiento (decrecimiento) de una de las variables se acompaña del crecimiento (decrecimiento) de la otra, e inversa en caso contrario. Hemos añadido, como en la tabla anterior, las categorías independencia y descripción verbal con objeto de una mejor comparación en el total de las tareas que se presentan en los textos. En la categoría "otras" se contabilizan siete situaciones, tres de ellas referidas a datos que representan una circunferencia ([L3], p. 232), una donde la pendiente de la función lineal es nula ([L3], p. 229), y tres que presentan relación cuadrática ([L7], p. 258; 260).

En la Tabla 2 observamos que la mayoría de situaciones planteadas corresponden a dependencia directa (entre el 46% y 85%), Le sigue en frecuencia las situaciones de relación inversa, y aquellas en que no es posible determinar su sentido. Destacamos los textos [L5], [L7] y [L8] por ser los que más variedad de situaciones proponen al respecto. Los resultados coinciden con los de Sánchez Cobo (1999). Como el autor, pensamos que es importante una distribución más equilibrada del sentido de la relación, para corregir (caso que se diese en los estudiantes) la concepción unidireccional de la asociación descrita por Estepa (1994) consistente en no aceptar la correlación inversa.

Tabla 2. Frecuencias (y porcentajes) de tareas según dirección de la relación

	L1	L2	L3	L4	L5	L6	L7	L8
Independencia	2(0,7)	5(2,3)	24(9,4)	12(5,3)	5(1,6)	6(3,4)	30(7,5)	21(7,1)
Directa	160(59,7)	188(85,1)	122(47,7)	153(68)	147(46,2)	109(61,9)	207(51,5)	166(56,5)
Inversa	60(22,4)	26(11,8)	72(28,1)	46(20,4)	115(36,2)	37(21)	93(23,1)	78(26,5)
Otras			4(1,6)				3(0,7)	
D. Verbal	46(17,2)	2(0,9)	34(13,3)	14(6,2)	51(16)	24(13,6)	69(17,2)	29(9,9)
Total	268	221	256	225	318	176	402	294

Contextos

Chevallard (1991) indica que en el proceso de transposición didáctica, una vez introducido un tema en el sistema de enseñanza, el dispositivo didáctico pretende, progresivamente, buscarle aplicaciones, que pueden no tener relación con aquellas en que originariamente se inició el concepto. La función que tienen es permitir finalmente la recontextualización del saber.

Con objeto de analizar este aspecto, también tenido en cuenta por Sánchez Cobo (1999), se han estudiado los contextos de aplicación en los libros de texto, encontrando una gran variedad, tanto para el desarrollo del tema, como para el planteamiento de tareas. Los hemos clasificado en seis categorías:

- *Fenómenos biológicos* (como la estatura de hijos – estatura de padres, precisamente el problema que históricamente dio origen a la idea de regresión);
- *Estudio en ciencias* (aumento de peso de un animal – mg diarios de un fármaco);
- *Deportivos* (distancia del jugador – número de encestes al jugar al baloncesto);
- *Economía* (consumo de energía per cápita y renta per cápita; kg de capturas de pescado y precio de subasta en la lonja);
- *Educativos* (notas de exámenes: física-matemáticas; matemáticas-filosofía);
- *Sociología y demografía* (renta per cápita - índice de natalidad);

Tabla 3 Contextos utilizados en los textos analizados

Contexto	L1	L2	L3	L4	L5	L6	L7	L8
Biológico	22(8,2)	24(10,9)	18(7)	10(4,4)	44(13,8)	31(17,6)	28(7)	36(12,2)
Ciencias	35(13,1)	73(33)	38(14,8)	30(13,3)	42(13,2)	29(16,5)	38(9,5)	27(9,2)
Deportes	12(4,5)		13(5,1)	5(2,2)	7(2,2)			
Economía	38(14,2)	18(8,1)	33(12,9)	12(5,3)	22(6,9)	26(14,8)	14(3,5)	28(9,5)
Educativos	28(10,4)	10(4,5)	5(2)	32(14,2)	54(17)	27(15,3)	23(5,7)	35(11,9)
Sociología	17(6,3)	38(17,2)	28(10,9)	32(14,2)	29(9,1)	22(12,5)	40(10)	29(9,9)
Sin contexto	116(43,3)	58(26,2)	121(47,3)	104(46,2)	120(37,7)	41(23,3)	259(64,4)	139(47,3)
Total	268	221	256	225	318	176	402	294

En la Tabla 3 se presentan los resultados del análisis realizado. Observamos un alto número de tareas descontextualizadas, en especial en [L7], superando a lo obtenido por Sánchez Cobo (1999) que obtuvo un 31,7% de tareas descontextualizados, aunque para relacionarlo con nuestro estudio se trataría de un 37,9% ya que su categoría “expresión matemática” es tratada en la presente investigación como “sin contexto“. Nuestro porcentaje suele ser mayor, siendo los textos [L2] y [L6] (26,2% y 23,3%, respectivamente) los que presentan menos tareas de este tipo. Esta alta presencia de tareas sin contexto es contraria a las recomendaciones actuales en la enseñanza, pues uno de los modos fundamentales de razonamiento estadístico, de acuerdo a Wild y Pfannkuch(1999) es la integración de la estadística con el contexto, y este tipo de tareas no facilitan este proceso.

El resto de contextos tienen una representatividad similar, siendo muy frecuente el contexto de ciencias, y escaso el contexto del deporte.

Representación gráfica

La representación gráfica alcanza un estatuto privilegiado en el estudio de la distribución bidimensional. En primer lugar, permite visualizar dicha distribución, facilitando su interpretación, y en segundo lugar constituye un recurso didáctico para la enseñanza posterior de la correlación y regresión, que se suelen introducir a partir de un diagrama de dispersión ya que permite visualizar fácilmente la intensidad y signo de la correlación y la tendencia lineal o no lineal. Se presentan a continuación las diferentes representaciones gráficas que encontramos en los textos analizados:

Diagrama de barras tridimensional. Es una representación espacial de la distribución bidimensional, cuyos valores se representan en un plano, utilizándose la tercera dimensión para representar la frecuencia de cada par de valores, con una barra de altura proporcional a su frecuencia.

Diagrama de dispersión y gráfico de burbujas. La representación gráfica más utilizada de la distribución bidimensional es el diagrama de dispersión o nube de puntos, donde cada par de valores se presenta en un sistema de coordenadas cartesianas. Cuando todos los datos poseen frecuencia absoluta igual a la unidad cada dato se corresponde con un punto en el plano. En otro caso, se sitúa dicho dato, y a su alrededor se dibujan tantos puntos como indique su frecuencia absoluta, aunque lo más indicado es optar por dibujar circunferencias con área proporcional a la frecuencia de cada par de valores (gráfico de burbujas). Como se ha indicado, las dos representaciones son muy útiles para interpretar la relación entre las variables de estudio, ya que permiten visualizar su intensidad (a través de la mayor o menor dispersión de la nube de puntos), su sentido (si la relación es directa o inversa) y el tipo (lineal o no), observando su tendencia (Sánchez Cobo, 1999). Además, el diagrama de burbujas es muy útil para visualizar en el plano, simultáneamente, hasta tres variables (el diámetro) o incluso cuatro, si mediante el color pudiera representarse una cuarta variable cualitativa.

Histograma tridimensional. Este tipo de representación se utiliza principalmente cuando las variables que conforman la variable estadística bidimensional son continuas. Permite visualizar en el espacio la distribución bidimensional, situando en el plano XY cada pareja de intervalos de clase, y levantando sobre el rectángulo resultante del cruce de estos intervalos un prisma de volumen proporcional a su frecuencia absoluta.

Pictograma tridimensional. Este gráfico permite visualizar la distribución de la variable estadística bidimensional en el espacio, a la vez que informa de su significado mediante la figura con la cual se representa. Es una variante del diagrama de barras e histograma tridimensional, ya

que cada barra o prisma se sustituye por una figura cuyo tamaño es proporcional a la frecuencia de los datos que representan.

En la Tabla 4 se muestra un resumen de la presencia de las representaciones gráficas en los textos analizados, mostrando su relevancia en el desarrollo teórico y/o práctico del tema.

Tabla 4. Presencia de representaciones gráficas en los textos analizados

	Presencia en el tema	L1	L2	L3	L4	L5	L6	L7	L8
Diagramas de dispersión	Desarrollo teórico y práctico	x	x	x	x	x	x	x	x
	Utiliza con frecuencias mayores a 1			x	x			x	x
Gráfico de barras	Utiliza		x	x	x				
	Presencia anecdótica	x						x	
Gráfico de burbujas	Utiliza							x	
	Presencia anecdótica	x		x					
Histograma	Utiliza			x					
	Presencia anecdótica	x						x	
Pictograma	Presencia anecdótica			x					

En cuanto a la representación de la variable bidimensional en el plano, podemos observar la destacada presencia del diagrama de dispersión en todos ellos, siendo en algunos la única representación gráfica utilizada ([L5], [L6], [L8])

Este tipo de representación no suele ir acompañada del gráfico de burbujas, donde menos de la mitad de los textos presentan este gráfico, y en su mayoría de modo anecdótico, con tan sólo una tarea de aplicación. Esta representación está especialmente indicada para distribuciones en que los datos posean frecuencia distinta a la unidad, que son las situaciones más comunes que encontramos en la realidad, no siendo muy habitual encontrar una muestra en la que todos los datos posean frecuencia uno. Por este motivo se incluyó, en el análisis de la presencia del diagrama de dispersión, el descriptor “Utiliza con frecuencias mayores a 1”, pues nos permite valorar si los textos contemplan este tipo de situaciones problemáticas en el tema, aunque no utilicen el gráfico de burbujas.

V. Conclusiones

Aunque el estudio de la competencia en la estimación de la correlación ha tenido poca relevancia en didáctica de la matemática, encontramos una amplia literatura en psicología que describe el razonamiento covariacional como una competencia fundamental para la toma de decisiones. El análisis de las estrategias intuitivas de los estudiantes en la estimación de la correlación a partir de diversas representaciones muestra, por un lado, la existencia de numerosas estrategias incorrectas; por otro, la de concepciones incorrectas y sesgos en la detección de la correlación, que a veces se mantienen tras la enseñanza (Batanero, Estepa y Godino, 1997; Batanero, Godino

y Estepa, 1998). Todo ello nos aporta información sobre posibles dificultades de los estudiantes, que los futuros profesores debieran conocer.

Algunos estudiantes muestran una concepción local, realizando su estimación a partir de sólo una parte de los datos. Moritz (2004) sugiere enseñar a los estudiantes a leer progresivamente punto a punto hasta una posterior generalización con los datos disponibles. Otra ayuda sería la utilización de un enfoque de variación temporal que permita a los estudiantes centrarse en el cambio de una variable a lo largo del tiempo para una posterior correspondencia entre variables no temporales.

Es también interesante alentar la expresión de las creencias y teorías previas, que serían poco a poco equilibradas mediante la información proporcionada por el estudio. Para ello conviene utilizar tareas que impliquen un razonamiento covariacional contra intuitivo, donde se cuestione de modo natural la fiabilidad del conjunto de datos de que se dispone, y para ello el profesor deberá prestar atención a los contextos con que diseña estas situaciones, con el fin de hacerlas emerger.

Por otro lado, Shaughnessy y Pfannkuch (2002), resaltan la importancia de situar la estadística en contexto y propiciar el proceso de transnumeración (Wild y Pfannkuch, 1999), donde el estudiante debe interpretar los datos y trabajar con ellos en diferentes representaciones para obtener conclusiones y hacer predicciones. En este sentido, se puede mostrar a los estudiantes algunos ejemplos reales como la paradoja de Simpson, que ha llevado a conclusiones equivocadas en la investigación. Un ejemplo, reciente (Saari, 2001) se produjo cuando en 1999, el Estado de California estableció un sistema de primas para los profesores de los colegios públicos en los que se lograra mejorar el rendimiento de los estudiantes. La evaluación de los centros mostró (con satisfacción de las autoridades educativas) que el rendimiento de los estudiantes había mejorado en la mayoría de los colegios, tanto para el grupo racial dominante, como para la minoría (hispanos). Sin embargo, al mezclar los datos de los dos grupos, el rendimiento global bajó en casi todos los centros, debido a una variación de la proporción de hispanos en los centros a lo largo del periodo.

Finalmente, estamos de acuerdo con Lavalle et al. (2006), en adelantar la enseñanza de la correlación, que podría comenzar entre el segundo y el tercer año de secundaria (14 a 15 años), cuando ya los estudiantes conocen los conceptos de función, función lineal, y poseen un dominio de la representación gráfica bidimensional. En principio, se podría hacer un tratamiento intuitivo, ampliando lo aprendido sobre distribuciones univariantes con el estudio descriptivo y la representación gráfica de datos bivariados. Luego se llevaría a cabo una aproximación a la correlación lineal a través del análisis de gráficos, para lo cual, los recursos tecnológicos como el uso de applets o la hoja de cálculo facilitarían mucho su determinación, para finalmente, incorporar los cálculos referidos al coeficiente de correlación y determinación.

Agradecimientos: Proyecto EDU2010-14947, FPI-BES-2011-044684 (MICINN-FEDER) y grupo FQM126 (Junta de Andalucía).