NÚMEROS
Revista de Didáctica de las Matemáticas
http://www.sinewton.org/numeros
ISSN: 1887-1984
Volumen 87, noviembre de 2014, páginas 25-35

El sentido de la correlación y regresión

María M. Gea, Carmen Batanero y Rafael Roa
(Universidad de Granada. España)

Fecha de recepción: 17 de febrero de 2014
Fecha de aceptación: 30 de julio de 2014

Resumen En este trabajo aplicamos nuestro modelo previo sobre sentido estadístico para proponer un modelo específico del sentido de la correlación y regresión, describiendo los componentes de la cultura y razonamiento estadístico específicos de este tema. Se muestran ejemplos de la forma en que dichos componentes se implementan en los textos de matemáticas de Bachillerato.

Palabras clave Sentido estadístico, correlación y regresión, libros de texto, Bachillerato.

Abstract We apply our previous model on statistical sense to propose a model for the sense of correlation and regression sense and describe the components of statistical literacy and reasoning which are specific for this topic. We present examples of how these components are implemented in high school mathematics textbooks.

Keywords Statistical sense; correlation and regression; textbooks; High school.

1. Introducción

El razonamiento covariacional es una actividad cognitiva fundamental en diversas actividades de la vida humana (Moritz, 2004; Zieffler, 2006; McKenzie y Mikkelsen, 2007) pues percibir, interpretar y predecir los sucesos que se presentan a lo largo de la vida depende de habilidades y destrezas para detectar covariaciones entre los acontecimientos (Alloy y Tabachnik, 1984). Sin embargo, la investigación previa sugiere pobres capacidades en los adultos para estimar la correlación, o realizar predicciones de una variable en función de otra, en ausencia de enseñanza específica sobre el tema. Asimismo se han descrito numerosas dificultades y errores conceptuales que continúan después de la enseñanza (Estepa, 2004; Gea, 2013). Una explicación es que el razonamiento humano en situaciones de incertidumbre está regido por valores y creencias del propio individuo. Debido a ello, el procesamiento de la información difiere de un proceso algorítmico, que produce una solución única para cualquier problema, dentro de una clase dada (Batanero, 2001).

Esta situación plantea a los profesores el reto de mejorar la enseñanza de la correlación y regresión, que matemáticamente recoge el estudio de la dependencia aleatoria y la modelización de datos bivariados, y que actualmente se introduce en primer curso de Bachillerato en las modalidades de *Ciencias y Tecnología* y *Humanidades y Ciencias Sociales* (MEC, 2007). En concreto, en esta segunda modalidad, y dentro del bloque 3 (*Probabilidad y estadística*), se incluye el siguiente contenido, muy similar en el Bachillerato de Ciencias y Tecnología:

> Distribuciones bidimensionales. Interpretación de fenómenos sociales y económicos en los que intervienen dos variables a partir de la representación gráfica de una nube de puntos. Grado de relación entre dos variables estadísticas. Regresión lineal. Extrapolación de resultados (MEC, 2007, p. 45475).

Igualmente se especifican los siguientes criterios de evaluación, similares en ambas modalidades de Bachillerato:

> Distinguir si la relación entre los elementos de un conjunto de datos de una distribución bidimensional es de carácter funcional o aleatorio e interpretar la posible relación entre variables utilizando el coeficiente de correlación y la recta de regresión.
>
> Se pretende comprobar la capacidad de apreciar el grado y tipo de relación existente entre dos variables, a partir de la información gráfica aportada por una nube de puntos; así como la competencia para extraer conclusiones apropiadas, asociando los parámetros relacionados con la correlación y la regresión con las situaciones y relaciones que miden. En este sentido, más importante que su mero cálculo es la interpretación del coeficiente de correlación y la recta de regresión en un contexto determinado (MEC, 2007, pp.45475-45476).

En este trabajo tratamos de desarrollar la idea de sentido estadístico, definida por Batanero, Díaz, Contreras y Roa (2013) como unión de la *cultura estadística* y el *razonamiento estadístico*, y aplicarla al caso específico de la correlación y regresión, con la finalidad de orientar la labor del profesor en el aula. Para ello, se parte del estudio previo de síntesis de la investigación didáctica sobre este tema, presentado en Gea (2013), complementado con un análisis del tema en los libros de texto de Bachillerato (Gea, Batanero, Contreras y Cañadas, 2013), y que ahora reinterpretamos para proponer un modelo de componentes del sentido de la correlación y regresión. Asimismo, se ejemplifican dichos componentes utilizando ejemplos de diferentes textos de Bachillerato, que actualmente se utilizan en el aula.

2. Componentes de la cultura sobre correlación y regresión

Un primer componente del sentido estadístico es la adecuada cultura estadística, que aúna el conocimiento básico sobre el tema, unido a unas actitudes positivas (Gal, 2002), que se particularizan para la correlación y regresión en lo que sigue.

2.1. Actitudes y creencias

En primer lugar se requiere unas disposiciones favorables, la superación de creencias erróneas sobre el tema, la valoración del método como instrumento de resolución de problemas y una actitud crítica ante el uso inapropiado o el abuso de la correlación y regresión. En concreto, algunos de los sesgos concretos que se deben evitar son los siguientes:

- La *correlación ilusoria* (Chapman, 1967) que consiste en dar prioridad a las propias creencias, frente a la relación entre dos variables, sin tener en cuenta la evidencia ofrecida por los datos. Ocasiona la percepción de la correlación cuando no existe, la sobreestimación de una correlación dada; e incluso la percepción de una correlación contraria a la existente.

- *Creencia en la transitividad de la correlación* (Castro-Sotos, Vanhoof, Van Den Noortgate y Onghena, 2009). Es decir, cuando una variable *X* está correlacionada con otra variable *Y*, y esta a su vez lo está con una tercera variable *Z*, pensar que *X* ha de estar correlacionada con *Z,* lo cual no siempre es cierto.
- *Juzgar la correlación entre dos variables X e Y sin controlar otras que las afectan* y que pueden cambiar el sentido o la intensidad de la correlación, efecto denominado paradoja de Simpson. Un ejemplo descrito por Saari (2001) es encontrar una mejora en el rendimiento académico de los alumnos, cuando se estudia este rendimiento centro a centro y luego descubrir que el rendimiento global no mejora al condensar todos los datos, debido a que no se tiene en cuenta la diferente proporción en cada centro de hijos de inmigrantes, que tienen dificultad con la lengua.
- *No considerar el efecto de regresión*. Es habitual que al correlacionar dos medidas sobre los mismos sujetos (por ejemplo, la estatura de una persona y la de su padre o la puntuación en dos pruebas sucesivas) las puntuaciones atípicas en la primera variable se acerquen al valor medio en la segunda. Galton descubrió este efecto que denominó reversión hacia la media y es debido a la normalidad de los datos bivariantes (Estepa, Gea, Cañadas, y Contreras, 2012). Sin embargo, algunas personas interpretan este efecto como un cambio; en el segundo ejemplo, se puede pensar en un efecto de aprendizaje en los estudiantes con puntuación excesivamente baja en la primera prueba.

2.2. Ideas fundamentales en correlación y regresión

Un adecuado sentido estadístico requiere, asimismo, el conocimiento de las ideas estadísticas fundamentales descritas por Burrill y Biehler (2011), que son las siguientes: datos, representación de la información y transnumeración, variabilidad aleatoria, distribución, asociación y ajuste de modelos entre dos variables, probabilidad, muestreo e inferencia. A continuación analizamos y desglosamos aquellas que aparecen en el estudio de la correlación y regresión en el nivel de enseñanza de Bachillerato:

Datos y Distribución. En el estudio de la correlación y regresión se manejan datos bivariantes, es decir, para cada individuo de una muestra se consideran conjuntamente los valores de las dos variables estadísticas que se pretende relacionar. Tendremos entonces *una variable estadística bidimensional,* formada por el conjunto de todos sus pares de valores posibles; si añadimos las correspondientes frecuencias conjuntas (o de aparición de cada par concreto de valores), se obtiene la correspondiente *distribución bidimensional*. Asociada a la misma, será necesario que los estudiantes diferencien las *frecuencias absolutas* y *relativas conjuntas*, *marginales* y *condicionales*. Sin embargo, en algunos textos de Bachillerato no se diferencian los conceptos de variable y de distribución bidimensional, y por lo general, las variables que forman la variable bidimensional se presentan con igual número de modalidades. Por otro lado, otros textos no presentan el estudio de las distribuciones condicionales y/o marginales (Gea, Batanero, Fernándes y Gómez, 2013).

Representación tabular y gráfica. Por su papel esencial en la organización, descripción y análisis de datos, las tablas y gráficos son un instrumento esencial en el análisis estadístico, y su conocimiento es parte de la cultura estadística (Arteaga, Batanero, Cañadas y Contreras, 2011). En el estudio de la correlación y regresión los datos se organizan en una tabla de doble entrada, cuyas celdas representan la frecuencia conjunta de los valores de la variable que se fijan en sus correspondientes filas y columnas. En los textos de Bachillerato se reconoce la importancia de esta forma de organizar la información, aunque la representación tabular más utilizada es el listado de datos (donde cada variable aparece en una columna), denominada en algunos textos *tabla de frecuencias bidimensional simple*. La representación gráfica más utilizada en Bachillerato es el diagrama de dispersión o nube de puntos, aunque también encontramos histogramas o gráficos tridimensionales y gráficos de burbuja

(Figura 1). Tanto el diagrama de dispersión como el gráfico de burbujas 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). El diagrama de burbujas permitiría también visualizar simultáneamente, hasta tres variables (representando la tercera mediante el diámetro) o incluso cuatro, si mediante el color pudiera representarse una cuarta variable.

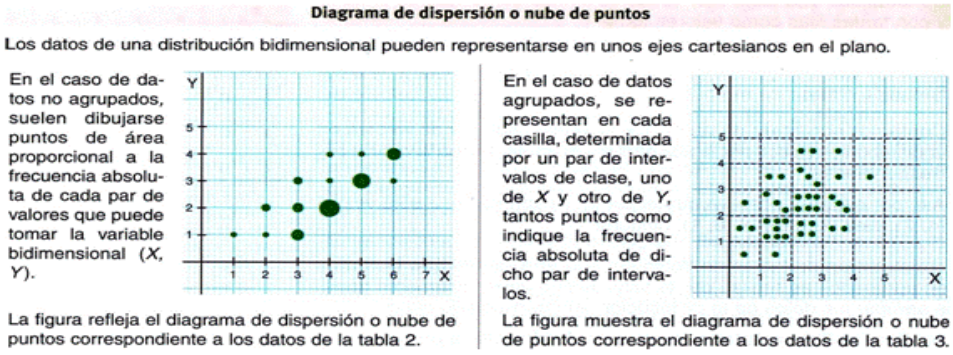


Figura 1. Diagrama de burbujas y diagrama de dispersión (Biosca et al., 2008, p.270).

Variabilidad aleatoria. La variabilidad se manifiesta en este tema mediante la dispersión de la nube de puntos respecto al modelo de regresión; que en Bachillerato se reduce al análisis de la mayor o menor desviación de los puntos a la recta de regresión. De hecho, es posible interpretar la línea de regresión como tendencia, y la distancia de los puntos a la misma como dispersión o variabilidad. Metafóricamente, se podrían usar los términos "señal" (para la recta) y "ruido" (la distancia de los puntos a la recta) o bien "estructura" y "residuos" (Engel y Sedlmeier, 2011). Así es que, el modelo de regresión tendrá como principal finalidad la predicción de una de las variables en función de la otra, y la evaluación de la variabilidad latente en los datos. Posteriormente esta "variabilidad" de los puntos alrededor del modelo se medirá mediante la covarianza y el coeficiente de correlación (que también miden la dirección de la correlación):

> El valor de la covarianza indica cómo se apartan a la vez las dos coordenadas de un dato respecto de la media. Si el resultado es positivo, quiere decir que los productos son positivos, esto es, los valores de x_i y de y_i se alejan en el mismo sentido de sus respectivas medias. Y, al contrario, si el resultado es negativo. En caso de que el valor sea cero o próximo a cero, la covarianza informa de que no hay relación entre ambas variables (Vizmanos et al, 2008a, p.251).

De hecho, el cuadrado del coeficiente de correlación o *coeficiente de determinación,* además de medir la bondad del ajuste de los datos al modelo, también puede interpretarse como una medida de variabilidad: La proporción de la varianza de la variable dependiente *Y* explicada por el modelo de regresión, como explican algunos textos:

> En ocasiones, con el fin de calcular la calidad o bondad del ajuste realizado mediante la recta de regresión y, por tanto, la fiabilidad de las predicciones que con ella se puedan realizar, se utiliza la expresión $(r^2 \cdot 100)\%$, que nos da el porcentaje en el que la variable *Y* se justifica por el valor de la variable *X*. (Bescós y Pena, 2008, p.185).

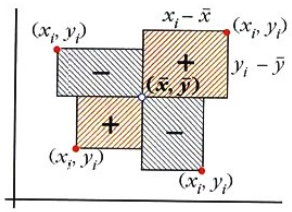
Dependencia funcional, aleatoria e independencia. Mientras que en una dependencia funcional a cada valor de una variable *X* (independiente) corresponde un solo valor de otra variable *Y*

(dependiente), en la dependencia aleatoria a cada valor de X corresponde una distribución de valores de Y, por lo que este concepto amplía el de dependencia funcional. Los libros de texto suelen resaltar esta diferencia, como en el ejemplo siguiente:

> En ocasiones se observa que existe una relación entre las variables, pero dicha relación no puede expresarse como una función matemática. En este caso se dice que entre las variables X e Y existe una dependencia estadística, que podrá ser fuerte o débil. (Monteagudo y Paz, 2008a, p. 338).

En este sentido, son muchos los textos que definen la idea de *independencia* como en el siguiente ejemplo: "*Cuando no existe dependencia funcional ni estadística, se dice que hay independencia estadística entre las variables*." (Monteagudo y Paz, 2008b, p.224). Esta definición es importante, pues la comprensión de la idea de independencia es base de muchos temas estadísticos posteriores; por ejemplo, en inferencia, un supuesto básico de aplicación de la mayor parte de contrastes estadísticos es admitir la independencia estadística de los datos de la muestra.

Covarianza y correlación. Como se ha indicado, con objeto de medir el signo y la intensidad de la dependencia entre dos variables estadísticas, se introducen otros dos conceptos importantes: la covarianza y la correlación. La covarianza permite analizar el signo de la correlación y se define formalmente en la mayoría de los textos, aunque por lo general, se acompaña de ejemplos que faciliten su comprensión. En este sentido, es de gran utilidad la explicación de su significado mediante la división en cuatro cuadrantes de la nube de puntos por las rectas correspondientes a las medias de cada variable (Figura 2), ya que permite al estudiante desarrollar una mejor comprensión. Este es el modo en que se razona el signo de la correlación y la covarianza en la propuesta didáctica de Holmes (2001). Así mismo, este análisis propicia que el estudiante comprenda más significativamente el cálculo del coeficiente de correlación en base al anterior.



Covarianza

Se llama **covarianza** al parámetro:

$$\sigma_{xy} = \frac{\Sigma(x_i - \bar{x})\,(y_i - \bar{y})}{n} = \frac{\Sigma\, x_i y_i}{n} - \bar{x}\,\bar{y}$$

Ambas expresiones, como es lógico, coinciden. La segunda de ellas es más cómoda para obtener numéricamente la covarianza.

En la figura adjunta, cada sumando $(x_i - \bar{x})\,(y_i - \bar{y})$ de la covarianza es el área de un rectángulo como los que aparecen en la figura.

Según donde esté situado (x_i, y_i) respecto a $(\bar{x}, \bar{y})$, el área $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ será positiva (rojo) o negativa (gris). Si los puntos están próximos a una recta de pendiente positiva, los sumandos son casi todos positivos y la covarianza es grande.

Figura 2. Definición e interpretación de la covarianza (Colera et al., 2008, p. 228).

Regresión. Encontrada una relación moderada o fuerte entre las variables, el siguiente paso es tratar de deducir un modelo matemático que permita predecir una de ellas en función de la otra. El concepto de regresión suele tratarse de modo implícito en los textos de Bachillerato, con algunas excepciones como en el siguiente ejemplo, donde se resalta su utilidad predictiva:

> La regresión consiste en tratar de encontrar una función matemática que relacione las variables X e Y de una distribución bidimensional, de forma que, si se conoce el valor de una variable, se puede calcular el correspondiente de la otra, con mayor o menor aproximación. (Monteagudo y Paz, 2008a, p. 340).

La diferencia entre *variable dependiente e independiente* ocupa un lugar central en el análisis de

la regresión, ya que, una vez aceptada la dependencia entre las variables del estudio, y con objeto de expresar en forma de ecuación o modelo una variable en función de otra, se necesita seleccionar qué variable servirá como dependiente o predictora. La definición explícita de variable dependiente e independiente se encuentra en pocos textos; un ejemplo es el siguiente: "*La variable dependiente es aquella que se quiere estimar, y la variable que se utiliza para ello se denomina variable independiente.*" (Vizmanos y cols., 2008a, p.254). En el resto se suele incluir implícitamente, pues se hace explícita la existencia de dos rectas de regresión diferentes:

> Si *X* se considera la variable independiente e *Y* la variable dependiente, la ecuación de la recta de regresión es: $y-\bar{y}=\frac{S_{xy}}{S_x^2}\left(x-\bar{x}\right)$. Esta recta se denomina recta de regresión de *Y* sobre *X*. A partir de ella, conocidos los valores de *X,* y sustituyéndolos en la ecuación, se pueden calcular con una cierta aproximación los valores de *Y*.
>
> Si se considera *Y* como variable independiente y *X* como variable dependiente, se obtiene la recta de regresión de X sobre Y, cuya ecuación es: $x-\bar{x}=\frac{S_{xy}}{S_y^2}\left(y-\bar{y}\right)$. Igual que en el caso anterior, conocidos los valores de *Y*, y sustituyéndolos en la ecuación de la recta, se obtienen con cierta aproximación los valores de *X*. (Monteagudo y Paz, 2008b, p.226).

Modelos de regresión y sus parámetros. Como señala Moore (2005), la regresión es un modelo general para comprender las relaciones entre variables, por lo que el concepto de modelo es fundamental en el tratamiento de la regresión. En muchos textos de Bachillerato, esta idea queda implícita, y se restringe al modelo lineal, con algunas excepciones, como por ejemplo:

> Si en una variable *(X,Y)* existe una correlación fuerte entre las variables *X* e *Y*, el análisis de la regresión permite encontrar la ecuación de la función matemática que mejor se ajusta a la nube de puntos. Esta puede ser una recta, una parábola, una exponencial, una cúbica, etc. (Monteagudo y Paz, 2008b, p.226).

En otros textos se aclara que el *método de mínimos cuadrados* permite obtener aquella recta que minimiza los cuadrados de las diferencias entre los datos teóricos y los reales. Y en algunos casos, se indican los parámetros que definen la recta, centrando el interés en la pendiente de la recta, para lo cual se diferencian los dos *coeficientes de regresión,* dependiendo de qué variable se considere como dependiente o independiente. En algunos casos, sin embargo, sólo se define el coeficiente de regresión de *Y* sobre *X,* como por ejemplo:

> La recta que hace mínima la suma $\sum d_i^2$ tiene por ecuación: $y=\bar{y}+\frac{\sigma_{xy}}{\sigma_x^2}\left(x-\bar{x}\right)$ se llama recta de regresión de *Y* sobre *X*. A la pendiente, $\frac{\sigma_{xy}}{\sigma_x^2}$, se la llama coeficiente de regresión (Colera et al., 2008, p. 230).

En Bachillerato raramente se alude al tratamiento de datos atípicos, aunque encontramos un texto en el que se presenta el procedimiento de cálculo de la recta de Tukey para obtener la recta de regresión respecto a la mediana (Vizmanos et al., 2008a; 2008b).

Estimación y bondad de ajuste. En cuanto a la utilidad de la recta de regresión, para realizar estimaciones del valor de la variable dependiente en función de la independiente hemos constatado que sólo algunos textos la resaltan, al igual que en el estudio de Sánchez Cobo (1999). Todavía menos analizan la bondad del ajuste realizado, definiendo el *coeficiente de determinació*n, aunque a veces esta idea queda implícita, como en el ejemplo ya citado de Bescós y Pena (2008).

3. Pensamiento y razonamiento estadístico sobre datos bivariados

Además de la comprensión de las ideas fundamentales que se han descrito en el apartado anterior, será también necesario desarrollar en los estudiantes el pensamiento y razonamiento estadístico en torno a la correlación y regresión. En primer lugar, sería necesario mejorar su estimación de la correlación a partir de diversas representaciones de datos, pues la investigación previa muestra sesgos en dicha estimación. Para ello, será necesario tener en cuenta las variables que influyen en la dificultad de la estimación de la correlación, que son las siguientes (Sánchez Cobo, Estepa y Batanero, 2000):

- *El signo de la correlación*, que puede ser positivo (dependencia directa), negativo (dependencia inversa) o nulo (independencia), siempre en caso de correlación lineal. 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 la correlación positiva, llegando en ocasiones a suponer que la correlación negativa es muy próxima a cero. En este sentido, en nuestro estudio previo (Gea, Batanero, Contreras y Cañadas, 2013) se observó que en algunos libros de texto hay muy pocos ejemplos de correlaciones negativas, llegando en algunos casos sólo al 12% de todos los ejemplos y actividades propuestas.
- *Intensidad de la dependencia*, siendo más fácil para los estudiantes percibir una correlación fuerte. De hecho, los estudiantes presentan dificultades al comparar diferentes valores del coeficiente de correlación (Sánchez Cobo, 1999). También este aspecto se puede mejorar en los libros de texto, que presentan en una amplia mayoría de situaciones con correlación muy próxima a 1, que escasamente aparecen en las aplicaciones reales, sobre todo en las ciencias sociales.
- *Concordancia entre los datos y las teorías previas* sugeridas por el contexto. Como se ha indicado al hablar de los sesgos en el razonamiento correlacional, muchos sujetos se guían preferentemente por sus teorías (en vez de usar los datos) cuando estiman una correlación. Por ello, es más fiable la estimación de la correlación cuando hay concordancia entre los datos y estas teorías previas que en caso contrario. Usualmente, en los ejercicios de los libros de texto las teorías previas y datos coinciden, por lo que sería interesante presentar algún ejemplo en que esto no ocurra.
- *El número de datos.* Respecto a la tarea de estimar un valor del coeficiente de correlación, Sánchez Cobo et al., (2000) indican que mejora cuando hay más datos. Por el contrario, los problemas tipo en los libros de texto presentan un número reducido de datos, pues son pocos los que están planteados para ser resueltos con ayuda de la tecnología.

El razonamiento sobre datos bivariados incluye la adquisición de estrategias adecuadas de análisis. Por supuesto, el cálculo e interpretación de la covarianza y correlación, principalmente mediante el uso de la tecnología para interpretar la correlación son las mejores estrategias. Pero el profesor también puede enseñar algunas estrategias intuitivas correctas basadas en el análisis del diagrama de dispersión. Dichas actividades se suelen incluir en los libros de texto, generalmente mediante ejemplos, como se muestra en la Figura 3:

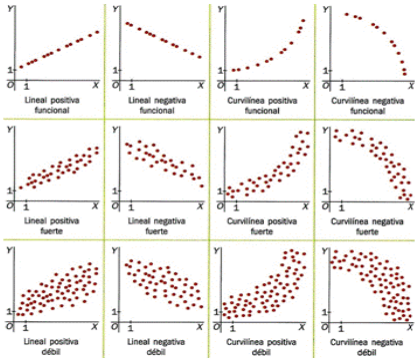


Figura 3. Interpretación del diagrama de dispersión (Vizmanos y cols., 2008b, p.322).

- *Comparación de la mayor o menor dispersión del diagrama de dispersión.* Se trataría de estimar la correlación entre las dos variables observando la dispersión de los puntos en torno a una línea de ajuste a los mismos. Esta estrategia es correcta, y permite estimar la intensidad de la correlación, pues a menor dispersión, aumenta la correlación. Los casos extremos serían cuando no existe dispersión (dependencia funcional) y el caso de independencia (máxima dispersión en la gráfica).
- *Crecimiento o decrecimiento.* Esta estrategia permite ver el signo de la correlación en caso de que la nube de puntos se ajuste a una función lineal. Consiste en estudiar la tendencia de la nube de puntos según la pendiente de la recta para justificar el sentido de la dependencia.
- *Comparación con un modelo matemático*. El estudiante compara la forma de la nube de puntos con una función conocida, por ejemplo, lineal o cuadrática. Si la forma de la nube se ajusta al modelo, deduce que hay correlación, y estima la intensidad de acuerdo al mayor o menor ajuste de los datos a éste. La estrategia funciona si la dispersión es pequeña y dependiendo del modelo elegido para comparar.

Sería también importante desarrollar la comprensión de la diferencia entre correlación y causalidad mediante la discusión con los estudiantes sobre la explicación de la correlación, que no siempre es debida a una relación causa-efecto. Otras posibles explicaciones sugeridas por Barbancho, (1973) son las siguientes:

- *Las variables pueden ser interdependientes* (cada variable afecta a la otra), como en el caso de la longitud de piernas y la altura de una persona;
- *La existencia de una tercera variable* que determine la correlación, esto es, las variables muestran dependencia pero es indirecta. Un ejemplo sería relacionar el índice de natalidad y la esperanza de vida, que presentan una correlación negativa. La proporción de mujeres que trabaja en un país afecta al producto nacional bruto, y al índice de natalidad (que disminuye) y con ello aumenta la esperanza de vida. Así es que estas dos variables están por ello correlacionadas, de modo indirecto, ya que su relación viene motivada por la influencia de otras variables como el número de mujeres empleadas en un país.
- La *concordancia* o coincidencia en preferencia u ordenación de una misma serie de datos, como por ejemplo, si dos profesores de modo independiente califican un mismo examen ya que, las calificaciones están correlacionadas pero la nota de uno no influye en la del otro.
- *Covariación casual o espúrea:* Cuando parece que en la covariación de dos variables hay cierta sincronía, lo cual podría interpretarse como la existencia de asociación entre ambas; sin embargo, ésta es casual o accidental.

En este sentido, algunos textos incluyen la definición de correlación espúrea o casual aproximándose a otras tipologías de covariación diferentes de la interdependencia y la dependencia

causal unilateral. Así por ejemplo, los siguientes textos muestran casos de covariación debida a terceras variables, y la denominan correlación espúrea:

> No siempre que dos variables generen una nube de puntos alargada existe correlación entre ellas. Son muchos los casos en los que dos caracteres varían a la vez, sin que por ello estén correlacionados. Por ejemplo: las canas y la miopía de las personas. Es posible que entre los canosos haya más miopes, pero no por ser canosos sino por ser mayores. Este tipo de falsas correlaciones se llaman espurias (Martínez, Cuadra y Heras, 2008, p. 252).
>
> Por ejemplo, es muy posible que exista una cierta correlación entre el número de restaurantes de una ciudad y el número de profesores que trabajan en ella. Esto se debe a que ambas variables están relacionadas con el número total de habitantes de la ciudad (Anguera et al., 2008, p.221).

4. Reflexiones finales

Un requisito necesario para desarrollar el sentido de la correlación y regresión en nuestros estudiantes es la preparación específica del profesor de matemáticas en este. Para ello habría que tener en cuenta en la formación de los profesores las seis facetas descritas por Godino (2009):

- *Faceta epistémica:* sería necesario desarrollar el conocimiento matemático del profesor, en nuestro caso de la correlación y regresión, y de los diferentes objetos matemáticos (problemas, lenguaje, conceptos, propiedades, argumentos y procedimientos) y procesos asociados al tema. Por ejemplo, el profesor debe saber identificar problemas de contextualización de la correlación y regresión o reconocer los medios de expresión matemática (diagramas de dispersión, notaciones, tablas,..) ligados al tema.
- *Faceta cognitiva:* implica el conocimiento del razonamiento, aprendizaje y dificultades de los estudiantes con la correlación y la regresión. Un ejemplo sería conocer si un estudiante está capacitado para resolver un problema o saber cómo ayudarlo; o bien reconocer el progreso en el aprendizaje del estudiante.
- *Afectiva:* conocimiento del grado de implicación (interés y motivación) del alumnado en el proceso de estudio. Ser capaz de buscar situaciones que motiven el interés de los estudiantes.
- *Mediacional:* uso de recursos tecnológicos, manipulativos y de todo tipo, apropiados para la enseñanza-aprendizaje del tema según el nivel de formación o el grado en que se imparte la enseñanza.
- *Interaccional:* conocimiento de modelos de comunicación entre los actores del proceso de instrucción; en particular se capaz de utilizar la evaluación para diagnosticar las dificultades de los estudiantes y favorecer la interacción entre ellos.
- *Ecológica:* conocimiento del grado en que el proceso de enseñanza y aprendizaje se ajusta al currículo y proyecto educativo del centro; establecer conexiones del tema con otras ideas matemáticas o de otras materias; estar abierto a la innovación docente y necesidades de la sociedad.

Finalmente, y como se indica en Batanero, Díaz, Contreras y Roa (2013), el profesor ha de ser también capaz de proponer a los estudiantes proyectos estadísticos adecuados a sus capacidades, pues los proyectos constituyen un recurso fundamental en el desarrollo del sentido estadístico de los estudiantes. El mismo proyecto descrito en el citado trabajo podría ser utilizado en la enseñanza de la correlación y regresión. Así mismo, remitimos al lector interesado en el estudio de estadísticas demográficas y su relación con indicadores económicos, al proyecto descrito en Batanero, Díaz y Gea (2011), en el que se tienen en cuenta las tareas correlacionales descritas en el presente trabajo, y que es adaptable a diversos niveles educativos.

Agradecimientos: Proyecto EDU2013-41141-P (MEC) y grupo FQM126 (Junta de Andalucía).

Bibliografía

Alloy, L. B. y Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, *91* (1), 112-149.

Anguera, J., Biosca, A., Espinet, M. J., Fandos, M.J., Gimeno, M. y Rey, J. (2008). *Matemáticas I aplicadas a las ciencias sociales*. Barcelona: Guadiel.

Arteaga, P., Batanero, C., Cañadas, G. y Contreras, J. M. (2011). Las tablas y gráficos estadísticos como objetos culturales. *Números,* 76, 55-67.

Barbancho, A. G. (1973). *Estadística elemental moderna*. Barcelona: Ariel.

Batanero, C. (2001). *Didáctica de la estadística*. Granada: Departamento de Didáctica de la Matemática.

Batanero, C., Díaz, C. y Gea, M. M. (2011). Estadísticas de la pobreza y desigualdad. En Batanero, C. y Díaz, C. (eds.) *Estadística con proyectos*, 97-124. Granada: Departamento de Didáctica de la Matemática.

Batanero, C., Díaz, C., Contreras, J. M. y Roa, R. (2013). El sentido estadístico y su desarrollo. *Números, 83,* 7-18.

Biosca, A., Doménech, M., Espinet, M. J., Fandos, M. J. y Jimeno, M. (2008). *Matemáticas I*. Barcelona: Guadiel.

Bescós, E. y Pena, Z. (2008). *Matemáticas aplicadas a las ciencias sociales*. Bilbao: Oxford University Press.

Burrill, G. y Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in trainingteachers. En Batanero, C., Burrill, G. y Reading, C. (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education - A joint ICMI/IASE study* (pp. 57-69). Dordrecht: Springer.

Castro-Sotos, A. E., Vanhoof, S., Van Den Noortgate, W. y Onghena, P. (2009). The transitivity misconception of Pearson's correlation coefficient. *Statistics Education Research Journal 8* (2), (pp. 33-55). Disponible en: www.stat.auckland.ac.nz/~iase/serj/.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, 6 (1), 151-155.

Colera, J., Oliveira, M. J., García, R. y Santaella, E. (2008). *Matemáticas aplicadas a las ciencias sociales I*. Madrid: Anaya.

Engel, J. y Sedlmeier (2011). Correlation and regression in the training of teachers. En Batanero, C., Burrill, G. F. y Reading, C. (Eds.), *Teaching statistics in school mathematics- challenges for teaching and teacher education. A joint ICMI/IASE study*, 247-258. Dordrecht: Springer.

Estepa, A. (2004). Investigación en educación estadística. La asociación estadística. En Luengo R. (Ed.), *Líneas de investigación en educación matemática* (pp. 227-255). Badajoz: Servicio de Publicaciones. Universidad de Extremadura.

Estepa, A., Gea, M. M., Cañadas, G. R. y Contreras, J. M. (2012). Algunas notas históricas sobre la correlación y regresión y su uso en el aula. *Números, 81*, 5-14.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M. y Scheaffer, R. (2007). *A curriculum framework for K-12 statistics education. GAISE report* Disponible en http://www.amstat.org/education/gaise/.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-25.

Gea, M. M. (2013). *Investigación didáctica en correlación y regresión*. Granada: La autora.

Gea, M. M., Batanero, C., Fernándes, J. A. y Gómez, E. (2013). Definiciones asociadas a la distribución de datos bidimensionales en textos españoles de Bachillerato. En Fernandes, J. A., Martinho, M. H., Tinoco, J. y Viseu, F. (orgs.) *Atas do XXIV Seminário de Investigaçao en Educaçao Matemática* (pp. 127-140). Braga: Centro de Investigaçao em Educaçao

Gea, M. M., Batanero, C., Contreras, J. M., y Cañadas, G. R. (2013) Variables y contextos en los problemas de correlación: Un estudio de libros de texto. *Trabajo presentado en el III EDEPA* (pendiente de publicación). Cartago: Instituto Tecnológico de Costa Rica.

Godino, J. D. (2009). Categorías de análisis de los conocimientos del profesor de matemáticas. *UNION 20,* 13-31.

Holmes, P. (2001). Correlation: From picture to formula. *Teaching Statistics*, 23 (3), 67-71.

Martínez, J. M., Cuadra, R., Heras, A. (2008). *Matemáticas aplicadas a las ciencias sociales. 1.º Bachillerato*. Madrid: McGraw-Hill.

McKenzie, C. R. M., y Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology,* 54 (1), 33-61.

MEC (2007). *Real Decreto 1467/2007, de 2 de noviembre, por el que se establece la estructura del bachillerato y se fijan sus enseñanzas mínimas*. Madrid: Autor.

Monteagudo, M. F. y Paz, J. (2008a). *1º Bachillerato. Matemáticas. Ciencias y tecnología*. Zaragoza: Edelvives (Editorial Luis Vives).

Monteagudo, M. F. y Paz, J. (2008b). *1º Bachillerato. Matemáticas aplicadas a las ciencias sociales*. Zaragoza: Luis Vives.

Moore, D. S. (1991). Teaching statistics as a respectable subject. En Gordon F. y Gordon S. (eds.) *Statistics for the twenty-first century*, 14-25. New York: Mathematical Association of America.

Moore, D. S. (2005). *Estadística aplicada básica*. Barcelona: Antoni Bosch.

Moritz, J. (2004). Reasoning about covariation. En Ben-Zvi, D. y Garfield J. (Eds.) ,*The challenge of developing statistical literacy, reasoning and thinking*, 221-255. Dordrecht: Kluwer.

Saari, D. (2001). *Decisions and elections. explaining the unexpected.* Cambridge: University Press

Sánchez Cobo, F. T. (1999). *Significado de la correlación y regresión para los estudiantes universitarios*. Tesis doctoral. Universidad de Granada.

Sánchez Cobo, F. T., Estepa, A. y Batanero, C. (2000). Un estudio experimental de la estimación de la correlación a partir de diferentes representaciones. *Enseñanza de las Ciencias*, *18* (2), 297-310.

Vizmanos, J. R., Hernández, J., Alcaide, F., Moreno, M. y Serrano, E. (2008a). *Matemáticas aplicadas a las Ciencias Sociales 1*. Madrid: SM.

Vizmanos, J. R., Hernández, J., Alcaide, F., Moreno, M. y Serrano, E. (2008b). *Matemáticas 1*. Madrid: Ediciones SM.

Zieffler, A. S. (2006). *A longitudinal investigation of the development of college students' reasoning about bivariate data during an introductory statistics course*. Tesis doctoral. Universidad de Minnesota.

María M. Gea Serrano, Licenciada en Matemáticas y Estadística, Máster en Estadística Aplicada, Máster en Didáctica de las Matemáticas por la Universidad de Granada y actualmente desarrolla su investigación en Didáctica de la Estadística en el Programa de doctorado en la Facultad de Ciencias de la Educación, Universidad de Granada, donde ejerce como profesora.
Email: mmgea@ugr.es

Carmen Batanero Bernabeu, Facultad de Ciencias de la Educación, Universidad de Granada. Fue miembro del Comité Ejecutivo de ICMI (International Comisión on Mathematical Instruction y Presidenta de IASE (International Association for Statistical Education). Ha coordinado varios congresos y proyectos de educación estadística.
Email: batanero@ugr.es

Rafael Roa Guzmán, es Doctor en Didáctica de la matemática y profesor Titular de Universidad de la Universidad de Granada. Ha participado en varios proyectos de investigación en educación estadística y ha publicado artículos y comunicaciones en congresos sobre esta temática.
Email: rroa@ugr.es