EVALUACIÓN DE LA COMPRENSIÓN DE LA CORRELACIÓN Y REGRESIÓN A PARTIR DE LA RESOLUCIÓN DE PROBLEMAS

ANTONIO ESTEPA
Universidad de Jaén, Spain
aestepa@ujaen.es

FRANCISCO T. SÁNCHEZ-COBO
Universidad de Jaén, Spain
fsanchez@ujaen.es

RESUMEN

En este trabajo presentamos un estudio exploratorio de caracterización del conocimiento que los estudiantes universitarios tienen sobre la correlación y regresión. Analizamos las soluciones a dos problemas de una muestra intencional de 193 estudiantes, que habían recibido un curso de estadística descriptiva en el ámbito universitario. Estudiamos los procedimientos de los estudiantes y discutimos las dificultades y errores que muestran los alumnos sobre el centro de gravedad del diagrama de dispersión, las rectas de regresión, el coeficiente de correlación, el tipo de relación entre las variables y la predicción.

Palabras Clave: Investigación en educación estadística; Aprendizaje; Evaluación; Correlación; Regresión

SUMMARY

In this paper we present an exploratory study intended to characterise University students' understanding of correlation and regression. We analyse the solutions to two problems from an intentional sample of 193 students who had previously received a course of descriptive statistics at the University. We study the student's procedures and discuss their difficulties and errors concerning the centre of gravity in the scatter plot, regression lines, correlation coefficient, type of relation between the variables and prediction.

Keywords: Statistics education research; Learning; Assessment; Correlation; Regression

1. INTRODUCCIÓN

Uno de los contenidos estadísticos de mayor importancia, tanto por su significación práctica como por su carácter instrumental para otros conceptos estadísticos, es la asociación estadística, entendiendo por tal la extensión del concepto de correlación a variables cualesquiera incluso no numéricas (Hildebrand, Lang y Rosenthal, 1977). La asociación

estadística aparece en los currículos de enseñanza secundaria y en los cursos universitarios, bien en sus tres campos de problemas (tablas de contingencia, correlación y comparación de muestras), bien, en alguno de ellos. La correlación y regresión es un tema de estudio en los cursos de iniciación a la estadística de los primeros niveles de muchas carreras universitarias.

En el presente trabajo presentamos un estudio exploratorio en que se profundiza en el estudio de la comprensión de conceptos relacionados con la correlación y la regresión en situaciones de resolución de problemas, en los primeros cursos universitarios. De esta manera caracterizaremos la comprensión de los estudiantes de la muestra sobre estos conceptos, después del proceso de estudio seguido en un curso académico. Los resultados obtenidos son de utilidad para futuras planificaciones de la enseñanza de un tópico tan importante en estadística como la correlación y regresión.

2. INVESTIGACIONES PREVIAS

A pesar de la relevancia estadística y curricular de la asociación estadística, la investigación llevada a cabo sobre este tópico es escasa en general y, procede de la Psicología y de la Didáctica de la Matemática, siendo más numerosas las investigaciones psicológicas que las didácticas. A continuación, vamos a presentar un resumen de las aportaciones que la investigación ha proporcionado sobre la correlación y regresión.

Las investigaciones psicológicas sobre correlación de variables continuas están orientadas al estudio de la capacidad de estimación de la correlación existente por parte de los sujetos. Dicha estimación se lleva a cabo por las personas a partir de su impresión subjetiva (no se hace uso de cálculo formal) de la dirección e intensidad de la relación existente entre las variables, en una tarea donde los datos se presentan en un diagrama de dispersión o tabla de valores. Aunque estos trabajos no se relacionan de manera directa con el nuestro, sí nos alertan de que las percepciones subjetivas de la correlación pueden influir en las respuestas dadas por los alumnos en la resolución de problemas. Las principales conclusiones de estos trabajos son las siguientes:

1. La estimación de la correlación es más exacta cuando se dan los datos en forma gráfica, que cuando se dan en forma tabular (Lane, Anderson y Kellam, 1985).
2. La estimación de la correlación mejora cuando hay más datos y la correlación es alta (Erlick y Mills, 1967).
3. Si existen teorías previas sobre la intensidad y signo de la correlación, se sobrestima la correlación real existente, siempre que ésta coincida con las ideas previas, mientras que, en caso contrario, es necesaria la presencia de una correlación fuerte para que los sujetos la detecten (Jennings, Amabile y Ross, 1982).

Las investigaciones de carácter didáctico se han llevado a cabo con el objeto de obtener resultados que mejoren la enseñanza aprendizaje de este importante tópico. Entre ellas destacaremos las que tienen una relación más directa con el presente trabajo.

Estepa y Batanero (1996) analizan los juicios intuitivos sobre la correlación que 213 estudiantes preuniversitarios sin instrucción previa dan a partir del diagrama de dispersión de variables con distinto tipo de correlación (directa, inversa, intensa, moderada). Clasifican las estrategias utilizadas para realizar los juicios y la influencia de las variables de tarea en las dichas estrategias. Identifican tres concepciones incorrectas de la correlación:

1. *Concepción determinista de la asociación:* Cuando el estudiante sólo admite un sólo valor para la variable dependiente para cada valor de la variable independiente, para

considerar que las variables están relacionadas. Estos estudiantes consideran independientes las variables si no se cumple esta condición.

2. *Concepción local de la asociación:* Cuando el estudiante realiza el juicio de asociación basándose únicamente en una parte de los datos presentados. Si la parte de datos considerada presenta un tipo de correlación, por ejemplo, directa, considera que éste es el tipo de correlación para todo el conjunto de datos.
3. *Concepción causal de la asociación*: Cuando el estudiante se basa únicamente en la existencia o no de relación causal para afirmar la existencia de correlación, sin tener en cuenta la correlación empírica observada en los datos.

En el estudio de Batanero, Estepa, Godino y Green (1996) sobre interpretación de tablas de contingencia, se encontró además, otra concepción incorrecta de la asociación estadística para las variables presentadas en tablas de contingencia; La *concepción unidireccional de la asociación* se presenta cuando el estudiante percibe la existencia de asociación, solamente cuando la relación entre las variables es directa, considerando la relación inversa como independencia.

Morris (1997) estudia las concepciones de los estudiantes sobre la correlación y sugiere las tareas más adecuadas para valorar su comprensión. Encuentra la concepción causal de la asociación y dificultades y confusiones que los sujetos exhiben con el signo de la misma. También preguntó a los estudiantes si encontraban la estadística difícil y útil para su formación, encontrando que era difícil para el 60% de la muestra estudiada y que ningún estudiante piensa que la estadística es irrelevante como materia de estudio. Nuestro trabajo completa el de Morris (1997), ya que añadimos el estudio de la regresión y estudiamos el interés que tiene el tópico correlación y regresión para los estudiantes, dentro de la asignatura de estadística.

Truran (1997) compara los datos de evaluación en dos cursos impartidos en universidades de Australia y Malasia sobre correlación y regresión. Estudia la interpretación que los estudiantes hacen del coeficientes de correlación y determinación, la pendiente y ordenada en el origen de la ecuación de regresión, la predicción y sus restricciones. Sus conclusiones reflejan pocas diferencias entre los estudiantes de ambos países. Casi todos los alumnos identifican correctamente la correlación moderada y negativa, aunque también encuentra la concepción determinista de la asociación. Asimismo, indica que algunos estudiantes argumentan, de modo razonable, las reservas que se deben tener con las extrapolaciones como son la intensidad de la correlación y el tamaño de la muestra. Este trabajo será una referencia para el nuestro.

Estepa y Sánchez-Cobo (1998) estudian la presentación de la correlación y regresión en los libros de texto de bachillerato, examinando, además, los ejercicios propuestos. Entre los hechos encontrados debemos subrayar el enfoque teoría-práctica en el desarrollo de los temas y la aplastante mayoría de ejemplos y ejercicios con conjuntos de datos con correlaciones positivas, frente a las negativas e independencia. Esto puede llevar a los estudiantes a considerar que sólo existe la correlación positiva (concepción unidireccional) o que para que exista correlación, ésta debe ser muy fuerte. Además, en los libros de texto de Educación Secundaria se presta poca atención al centro de gravedad de los diagramas de dispersión$(\bar{x}, \bar{y})$, es decir, el punto definido por las medias de las dos variables, a la distinción entre la variable explicativa y la explicada y a los problemas de predicción, utilizando las rectas de regresión. Este estudio nos proporcionó la base para diseñar la presente investigación.

3. METODOLOGÍA DE LA INVESTIGACIÓN

3.1. OBJETIVO DEL ESTUDIO

Estos resultados nos motivaron a profundizar en el conocimiento de la comprensión de los conceptos de correlación y regresión por parte de los estudiantes. El presente artículo reanaliza parte de los datos obtenidos por Sánchez-Cobo (1999), sobre la comprensión de la correlación y regresión en estudiantes universitarios. En el citado trabajo se estudió la estimación de la correlación que hacen los alumnos a partir de diversas representaciones (descripción verbal, diagrama de dispersión, tabla y coeficiente de correlación), observando que la exactitud de la estimación depende del tipo de representación (tabla, gráfico o expresión verbal), intensidad y signo de la correlación y tipo de covariación (Sánchez-Cobo, Estepa y Batanero, 2000). Este tipo de tarea es también explorada por Moritz (2002) con alumnos de educación primaria, llegando a la conclusión de que es importante tener en cuenta los procesos de traslación entre estas diferentes representaciones.

También completamos la investigación de Estepa y Batanero (1996), cuyos estudiantes no habían estudiado el tema de correlación y regresión y a los que se les pedía que estimasen la relación existente entre dos variables en forma intuitiva, guiándose tan sólo del diagrama de dispersión. Nuestra muestra de estudiantes ha seguido un curso de iniciación a la estadística, estudiando el tema de correlación y regresión y se les ha pedido calcular el coeficiente de correlación para interpretarlo y decidir que tipo de relación presenta el conjunto de datos del problema propuesto.

Los conceptos estudiados a continuación son: la relación existente entre el centro de gravedad de la nube de puntos $(\bar{x}, \bar{y})$ y la recta de regresión, el cálculo e interpretación del coeficiente de correlación, el tipo de relación entre las variables, determinación de las rectas de regresión, interpolación y discriminación entre las dos rectas de regresión. El estudio está basado en el análisis de los procedimientos, aciertos, errores y dificultades que los estudiantes manifiestan en sus respuestas a las cuestiones de los problemas planteados. Identificamos los procedimientos de resolución utilizados por los estudiantes, estudiando su adecuación y corrección en su desarrollo. Con esto completamos, en parte, las investigaciones anteriores aportando conocimientos válidos para la enseñanza del tema.

3.2. PROBLEMAS PLANTEADOS

En lo que sigue se estudian y discuten los resultados de dos problemas, que formaron parte de un cuestionario más amplio completado por los alumnos de la muestra, parte de los cuales fueron publicados en Estepa y Sánchez-Cobo (2001). Estos problemas que analizaremos a continuación se presentan en forma de prueba de ensayo, con respuestas abiertas y, en consecuencia, produce datos de tipo cualitativo.

Problema 1. Una recta de regresión tiene una pendiente de 16 y corta al eje de ordenadas en el punto y = 4. Si la media de la variable independiente es 8, ¿Cuál es la media de la variable dependiente?

Problema 2. La siguiente tabla muestra el número de bacterias por unidad de volumen que está presente en un cultivo después de un cierto número de horas.

Número de horas	1	2	3	4	5
Número de bacterias por unidad de volumen	18	21	33	54	61

a) Calcule el coeficiente de correlación lineal.

b) Indique qué tipo de relación (directa, inversa o independencia) existe entre ambas variables.

c) Determine la recta de regresión de y, número de bacterias por unidad de volumen, sobre x, número de horas.

d) ¿Qué número de bacterias cabe esperar que habrá, transcurridas 2.5 horas?

e) ¿Qué tiempo deberá pasar para que el número de bacterias del cultivo sea 27?

El primer problema fue tomado de Cruise, Dudley y Thayer (1984, pp. 288). Para su resolución los estudiantes deben utilizar los siguientes conceptos matemáticos: recta, pendiente, ordenada en el origen. Respecto a los conceptos estadísticos, son necesarios recordar el hecho de que la recta de regresión pasa por el centro de gravedad del diagrama de dispersión, fundamental para la aproximación intuitiva de la recta de regresión a partir del diagrama de dispersión. De esta manera se relaciona el contexto geométrico y el estadístico de esta situación (Truran, 1997). Por último, el estudiante debe usar la predicción, esencial en el estudio de la regresión.

El segundo problema es una adaptación del que aparece en Vizmanos y Anzola (1988, pp. 372). Además de los cálculos del coeficiente de correlación y de las rectas de regresión, se pretende caracterizar el uso que los alumnos hacen de ellos. En el apartado b) se busca que el estudiante interprete el coeficiente de correlación, ya que algunos autores (Morris, 1997; Truran, 1997) sugieren dificultades en esta tarea. Los apartados d) y e) se dirigen a explorar si el alumno ha interiorizado debidamente la predicción y si llega a comprender su naturaleza estocástica (Truran, 1997), haciendo un uso conveniente de ella, es decir, empleando la recta de regresión adecuada y haciendo referencia al valor promedio.

Antes de decidir el cuestionario definitivo, una primera versión del mismo fue ensayada en una muestra piloto, compuesta por un reducido número de alumnos, distintos de los que participaron en la muestra final, con el fin de estudiar su legibilidad y facilidad de comprensión. Teniendo en cuenta los resultados obtenidos decidimos el cuestionario definitivo.

3.3. MUESTRA

En los centros educativos, los alumnos están agrupados en tipos de estudios y niveles de enseñanza, y en consecuencia, cuando se plantea una investigación educativa, es difícil conseguir una muestra aleatoria. Por tanto, muchas veces, las muestras están constituidas por grupos naturales de estudiantes que cursan una asignatura de una carrera determinada, y son intencionales, como en nuestro caso, constituyendo la investigación un cuasi-experimento, en el sentido de Cook y Campbell (1979). Al ser nuestra muestra de carácter intencional, somos conscientes de las limitaciones en la generalización de nuestros resultados.

La muestra final estuvo formada por 193 estudiantes de la Universidad de Jaén, 57 hombres (30%) y 136 mujeres (70%), con una edad media de 20 años. De ellos, 104 (53.8%) estudiaban la Diplomatura de Empresariales y 89 (42.2%) la Diplomatura de Enfermería; 109 (56.4%) estudiantes tenían una formación preuniversitaria orientada a ciencias y 84 (43.6%) a letras. Por otra parte, 117 (60.6%) estudiantes no habían cursado Estadística en su formación preuniversitaria. Además, 155 (80.3%) de los estudiantes indicaron que los temas de correlación y regresión tienen suficiente, bastante o mucho interés dentro del programa de la asignatura de estadística que estudiaban. Finalmente, 162 (83.9%) de los estudiantes sugirieron que la asignatura de Estadística tiene suficiente, bastante o mucho interés para su formación dentro del plan de estudios. Estos resultados son ligeramente inferiores a los de Morris (1997), ya que todos sus estudiantes encontraron relevante la asignatura de estadística.

El programa de la asignatura de Estadística que estudiaban los alumnos de nuestra muestra comprendía contenidos fundamentales de estadística descriptiva, tales como: tablas de frecuencias y gráficos, medidas de centralización y dispersión, simetría y apuntamiento, variables estadísticas bidimensionales, tablas de contingencia, covarianza y correlación, regresión lineal y polinómica, muestreo, distribución en el muestreo, intervalos de confianza y test de hipótesis. Los problemas que presentamos fueron elegidos teniendo en cuenta el análisis de la enseñanza realizada en la programación del profesor del tema de correlación y regresión y los apuntes de dos estudiantes.

A los 193 estudiantes de la muestra se les pidió que cumplimentaran un cuestionario completo, dándoles, antes de comenzar unas breves instrucciones verbales, que enfatizaban y aclaraban las expuestas por escrito en el cuestionario. Como se puede observar en la lectura de los problemas propuestos, las respuestas esperadas de los estudiantes son abiertas.

Para analizar las respuestas de los estudiantes en cada uno de los apartados de los problemas, se codificaron las respuestas en cada cuestión y se agruparon según su similitud, obteniendo las distintas categorías de respuestas con las que se llevó a cabo un análisis cualitativo (Miles y Huberman, 1984; Huberman y Miles, 1994). Finalizamos el análisis con la clasificación de las categorías de respuestas, según el procedimiento utilizado y la corrección de las mismas, obteniendo su frecuencia y porcentaje.

4. RESULTADOS Y DISCUSIÓN

4.1. CÁLCULO DE LA MEDIA DE LA VARIABLE EXPLICADA

En el problema 1 el estudiante ha de utilizar sus conocimientos analíticos sobre la regresión lineal, incluyendo el centro de gravedad $(\bar{x}, \bar{y})$, su relación con la recta de regresión, la pendiente y la ordenada en el origen (Truran, 1997). Asimismo, debe tenerse en cuenta que, mientras que la correlación es una relación simétrica, pues su valor no depende de cuál sea la variable elegida como explicativa en la variable estadística bidimensional, en el caso de la regresión, hay que tener en cuenta el tipo de covariación (Barbancho, 1973).

Tabla 1. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 1 según procedimiento de resolución

Procedimiento	Tipo de solución			Número de alumnos que usa el procedimiento
	Correcta	Incorrecta	No responde	
$y = mx + n$	31 (38.3)	50 (61.7)		81
$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$	8 (34.8)	15 (65.2)		23
Uso de las dos expresiones anteriores	2 (100)			2
	2 (22.2)	7 (77.8)		9
$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$		4 (100)		4
Uso de las dos rectas de regresión	2 (66.7)	1 (33.3)		3
Uso de un parámetro estadístico		7 (100)		7
No responde			64(100)	64
Total	45 (23.3)	84 (43.5)	64 (33.16)	193

Si la dependencia es causal unilateral, las variables explicativa y explicada quedan determinadas de forma unívoca. Si el tipo de covariación es uno de los cuatro restantes - interdependencia, dependencia indirecta, concordancia o covariación casual - el resolutor debe decidir qué recta de regresión hay que emplear, *y* sobre *x* o *x* sobre *y*, y, desde ese momento, quedan determinadas ambas variables, explicativa y explicada. Puesto que en nuestro caso no se propone un contexto, el estudiante puede elegir entre las dos rectas.

Este problema ha sido difícil para esta muestra de estudiantes, ya que han dado respuesta solamente 129 estudiantes (66.8% de la muestra), siendo correctas sólo 45 (34.9% de la muestra). En la Tabla 1, hemos clasificado las respuestas, por procedimiento seguido y tipo de solución. De los 129 estudiantes que han respondido, 122 (94.57% de respuestas) han utilizado una o ambas rectas de regresión para resolver el problema. La principal dificultad encontrada en la resolución de este problema es que algunos estudiantes no tienen en cuenta que el centro de gravedad $(\bar{x}, \bar{y})$ está contenido en la recta de regresión. En la enseñanza este tema sólo se trató de manera implícita, aunque se hizo referencia a tal punto, como punto de corte de las dos rectas de regresión. A continuación detallamos las soluciones.

Uso de la Recta de Regresión de Y sobre X

El 80.62% de los alumnos que han respondido han utilizado en sus cálculos la expresión de la recta de regresión de *y sobre x*. Distinguimos tres casos:

1. *Uso de la expresión explícita de la recta de regresión y = mx + n*, siendo *m* la pendiente y *n* la ordenada en el origen de la recta de regresión de *y* sobre *x* (81 sujetos). Las condiciones expresadas en el enunciado no determinan de forma unívoca las variables dependiente e independiente. En consecuencia, una de las dificultades más frecuentes ha sido confundir la variable explicativa con la explicada (27 alumnos de los 50 que han llegado a una solución incorrecta a partir de la ecuación *y = mx + n*). Nueve estudiantes han hecho una interpretación inadecuada de los parámetros *m* y *n* en dicha ecuación, coincidiendo con una dificultad ya señalada en Truran (1997), conmutando sus valores, o bien, asignándoles valores no apropiados. Un ejemplo de respuesta es: *"y = n +mx, y = 4, a = 16, x = ?, 4 = 16 - 8x, 12 = 8x, x = 12 / 8 = 1.5"*. Las demás dificultades no son significativas debido a su baja frecuencia.
2. *Uso de la expresión punto-pendiente de la recta de regresión* (23 alumnos han empleado la expresión $y - \bar{y} = (S_{xy} / S_x^2)(x - \bar{x})$). Una dificultad es que, al tener la necesidad de conocer un punto de la recta para determinar esta ecuación, no se han dado cuenta que éste podría ser el correspondiente a la ordenada en el origen. Otra dificultad ha sido el uso de una expresión inadecuada de la recta de regresión *y sobre x* (7 estudiantes). También vuelve a aparecer la confusión entre las variables dependiente e independiente (8 estudiantes). Con este procedimiento, sin embargo, ha sido más fácil para los estudiantes interpretar cuál es la pendiente de la recta (S_{xy} / S_x^2), que dar un valor adecuado a la pendiente *m,* cuando emplean la ecuación explícita de la recta (*y = mx+n*).
3. *Otros procedimientos.* Dos alumnos se sirven de las dos expresiones anteriores de la recta de regresión de *y* sobre *x*: la ecuación explícita y la ecuación punto-pendiente. Estos dos alumnos no recordaban que el centro de gravedad pertenece a la recta de regresión, pero observan que la media de la variable dependiente se encuentra en la expresión de la ecuación punto-pendiente, comparan ambas expresiones y determinan la media de la variable explicada, llegando a una solución correcta, como se muestra en la siguiente respuesta:

$$\text{“}y \text{ sobre } x\text{:}\ \ y = mx + n,\ \ y = 16x + 4,\ \ y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}),\ \ m = \frac{S_{xy}}{S_x^2} = 16$$

$$n = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x},\ \ n = \bar{y} - 16.8,\ \ \bar{y} = 132\text{”}$$

Uso de la Recta de Regresión de X sobre Y

Trece alumnos, (10.1 % de los 129 que han respondido este problema), han utilizado la recta de regresión de *X* sobre *Y*. Dado que el problema se plantea en forma abierta, no se puede considerar que confundan la variable dependiente con la independiente. De ellos, 9 emplean la expresión de la ecuación explícita de la recta de regresión, $x = m'y + n'$, de los cuales sólo 2 responden de forma correcta. Los 4 restantes usan la ecuación punto-pendiente de la recta de regresión, $x - \bar{x} = (S_{xy}/S_y^2)(y - \bar{y})$, repitiéndose las dificultades encontradas anteriormente: confusión entre la variable explicativa y la explicada, intercambio de papeles entre la ordenada en el origen y pendiente de la recta de regresión o intercambio de parámetros de la ecuación explícita de la recta de regresión. Un ejemplo de respuesta de esta última dificultad es el siguiente:

$$\text{“}n' = \bar{x} - m'\bar{y},\ \ 16 = 8 - 4\bar{y},\ \ \bar{y} = \frac{-16 + 8}{4} = 2\text{”}$$

Uso de las dos Rectas de Regresión Y sobre X y de X sobre Y

Tres alumnos se sirven de ambas rectas de regresión, *y* sobre *x* y *x* sobre *y*. Podemos considerar que ésta es la mejor respuesta, pues, como indicábamos anteriormente, el problema se planteó de forma abierta y, por tanto, es plausible considerar a *x* como variable explicativa o como variable explicada.

Síntesis de dificultades

En resumen, podemos concluir que la dificultad más importante con la que se enfrentan los estudiantes de la muestra al resolver el problema 1 es la de discriminar entre la variable explicativa y la variable explicada, lo que ocurre a 38 alumnos (50% de los estudiantes que responden de forma incorrecta). Aunque las expresiones de la ecuación de la recta de regresión, punto-pendiente y explícita, son, obviamente, equivalentes, estos alumnos emplean como herramienta más operativa la punto-pendiente, pues en ella identifican mejor la pendiente que en la forma explícita; sin embargo, utilizan con más frecuencia la forma explícita, probablemente debido al enunciado del problema.

Finalmente, hemos de advertir que 7 alumnos de 129 intentan resolver este problema apoyándose en conceptos estadísticos relacionados con los datos suministrados; para ello usan parámetros estadísticos como, por ejemplo, la media, la covarianza, etc., siendo aplicados todos estos procedimientos de forma inadecuada.

4.2. CÁLCULO DEL COEFICIENTE DE CORRELACIÓN

Las respuestas de 168 estudiantes (87.05% de la muestra) al apartado a) del problema 2 se resumen en la tabla 2. La enseñanza recibida proporcionaba dos fórmulas para calcular el coeficiente de correlación: $r = S_{xy}/(S_xS_y)$ y el coeficiente de determinación $r^2 = m \cdot m'$, siendo *m* y *m'* los coeficientes de regresión lineal de las rectas *y sobre x* y *x sobre y*. Los estudiantes

de esta muestra utilizan más frecuentemente la primera, ya que es más operativa.

Tabla 2. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 2.a según procedimiento de resolución

Procedimiento	Tipo de solución			Número de alumnos que usa el procedimiento
	Correcta	Incorrecta	No responde	
Fórmula $r = S_{xy} /(S_x S_y)$	106 (65.8)	55 (34.2)		161
Fórmula incorrecta de r		5 (100.0)		5
Coeficiente de determinación		2 (100.0)		2
No responde			25 (100)	25
Total	106 (54.9)	62 (32.1)	25 (13.00)	193

Aproximadamente un tercio de los 161 alumnos que han utilizado una fórmula correcta para calcular el coeficiente de correlación, han dado una respuesta inadecuada, resultados superiores a los de Morris (1997), quien sólo obtuvo un 20% de respuestas correctas, a pesar de dar la fórmula para calcular el coeficiente de correlación. Los principales errores se producen en los cálculos (22 de las 55 respuestas erróneas, lo que supone un 40 %, mientras que Morris obtuvo el 62.5% en su investigación); también ha influido un uso inadecuado de la calculadora o el uso de expresiones inadecuadas para el cálculo de la desviación típica y covarianza. Algunos alumnos dibujan un diagrama de dispersión para deducir el resultado pedido. Entre las argumentaciones ofrecidas es interesante reseñar las siguientes:

1. Tener conciencia de que el resultado obtenido no es correcto, ya que no está comprendido en el intervalo *[-1, 1],* pero no ser capaz de calcular el correcto. Una respuesta de este tipo es la siguiente: *"r = 2.24. Debe salir entre -1 ≤ r ≤ 1. Este dato debe estar mal"*.
2. Reconocer que el resultado obtenido es incorrecto a la vista del diagrama de dispersión. Una respuesta de este tipo es la siguiente: *"Debe estar mal porque se ve claramente que existe correlación"* (cuando observa el diagrama de dispersión que el propio estudiante ha construido).

Las dos respuestas anteriores sugieren que el alumno conoce el algoritmo adecuado para responder a la pregunta formulada, pero no es capaz de realizarlo correctamente. A la vista de estos resultados, podemos concluir que la mayoría de los alumnos utiliza una expresión adecuada para calcular el coeficiente de correlación, que ciertos errores son debidos al cálculo, coincidiendo con los resultados de Morris (1997) y son pocos los errores debidos a conocimiento incorrecto sobre la correlación.

4.3. DETERMINACIÓN DE LA RECTA DE REGRESIÓN DE *Y* SOBRE *X*

Las respuestas de los 162 alumnos que han contestado a esta cuestión (83.9% de la muestra) se presentan en la tabla 3, junto con sus procedimientos. Los estudiantes han calculado sólo una recta de regresión (82.1 % del total de respuestas), o las dos (17.9 % del total de respuestas). Es de resaltar que la mitad de los estudiantes que responden (81 de 162) lo hacen de forma inadecuada, siendo ésta una actividad tan primordial para este tema.

De los 133 alumnos que calculan una sola recta de regresión, 97.0% utilizan la expresión de la ecuación punto-pendiente de la recta de regresión de *y* sobre *x,* mientras que el resto (3.0%) usan la recta de regresión de *x* sobre *y.* El 17.9 % hallan tanto la recta de regresión de *y* sobre *x* como la de *x* sobre *y*. Esto puede ser debido a que no saben claramente qué recta hay que determinar en este apartado o a que, como posteriormente - Problema 2 apartado e) -

necesitarán la recta de regresión de *x* sobre *y*, deciden obtenerla en este apartado, pero no hacen ninguna indicación al respecto. En cuanto a los errores que cometen son, esencialmente, similares a los expuestos cuando hemos tratado el cálculo del coeficiente de correlación.

Tabla 3. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 2c según procedimiento de resolución

Procedimiento	Tipo de solución			Número de alumnos que usa el procedimiento
	Correcta	Incorrecta	No responde	
Calcula la recta de regresión de y sobre x	81 (62.8)	48 (37.2)		129
Calcula la recta de regresión de x sobre y		4 (100.0)		4
Calcula ambas rectas de regresión		29 (100.0)		29
No responde			31(100)	31
Total	81 (42.0)	81 (42.0)	31 (16.0)	193

4.4. INTERPRETACIÓN Y PREDICCIÓN

Interpretación de la Relación entre las Variables

Los procedimientos que han empleado los estudiantes de la muestra para llevar a cabo el juicio de asociación pedido en el problema 2, apartado b) se han agrupado en 6 categorías, cuyas frecuencias y respuestas obtenidas se ofrecen en la Tabla 4.

Tabla 4. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 2b según procedimiento de resolución

Procedimiento	Respuesta				Número de alumnos que usa el procedimiento
	Dependencia directa (correcto)	Dependencia inversa	Independencia	No interpreta	
Coeficiente de correlación	46 (90.2)	1 (2.0)	3 (5.8)	1 (2.0)	51
Variación conjunta	17 (94.4)		1 (5.6)		18
Covarianza	17 (94.4)	1 (5.6)			18
Coeficiente determinación	3 (50.0)	1 (16.7)	2 (33.3)		6
Otras	3 (42.9)		3 (42.9)	1 (14.2)	7
Sin estrategia	57 (61.3)	2 (2.2)	5 (5.4)	29 (31.1)	93
Total	143 (74.1)	5 (2.6)	14 (7.2)	31 (16.1)	193

- *Uso del coeficiente de correlación*. Si el cálculo e interpretación son correctos, la respuesta es correcta, como la del siguiente estudiante: *La relación es directa ya que el coeficiente de correlación es positivo*. Aunque 168 alumnos han calculado el coeficiente de correlación en el apartado a), menos de un tercio de ellos, lo utilizan explícitamente para interpretar la relación entre las variables. Un hecho estadístico destacado es que 'la magnitud (fuerza) del coeficiente de correlación es por completo independiente de su dirección (positiva o negativa)" (Phillips, 1992, pp. 113). Sin embargo, unos pocos estudiantes confunden la intensidad y el sentido de la correlación, como se ve en la respuesta: *Existe una relación directa ya que el coeficiente de correlación es alto*. Probablemente estas respuestas se deban a que en la enseñanza se pone mucho énfasis en la correlación como medida de la intensidad de la relación, del mismo modo que se hacía

en los libros de texto de bachillerato (Estepa y Sánchez-Cobo, 1998), lo que puede incidir en que los alumnos confundan magnitud con dirección de la correlación.

- *Variación conjunta de las variables.* Cuando el estudiante argumenta la existencia de relación al observar la variación paralela de las dos variables, como arguye este alumno: *Relación directa puesto que a medida que va pasando el tiempo, el número de bacterias aumenta*. Esta argumentación puede constituirse en un obstáculo, si el estudiante tiende a pensar que existe proporcionalidad entre las variables.
- *Uso de la covarianza*. Los estudiantes que usan la covarianza emplean el signo de ésta para decidir el sentido de la relación, como argumenta este alumno: *Es una relación directa ya que la covarianza es positiva.* Algunos estudiantes creen erróneamente que el coeficiente de correlación se utiliza para decidir si hay relación lineal y la covarianza para especificar el tipo de correlación existente, como manifiesta este estudiante: *A través del coeficiente de correlación se puede afirmar que existe una dependencia lineal alta. Para determinar si es directa o inversa se utiliza la covarianza* S_{xy} *=23.8. Al ser positiva será directa.*
- *Uso del coeficiente de determinación*. Algunos estudiantes usan la bondad del ajuste como criterio para juzgar la relación, por ejemplo: *La relación es directa e intensa ya que* r^2 *se aproxima a 1*. Aquí aparece nuevamente la confusión entre intensidad y dirección de la relación, un estudiante responde: *Relación inversa ya que el valor de* r^2 *= 0.4568 no está próximo a 1.*
- *Otras estrategias*. En esta categoría se incluyen diversos procedimientos y argumentaciones, como utilizar un intervalo de variación del coeficiente de correlación, usar el diagrama de dispersión o comparar la asociación y la proporcionalidad.
- *Sin estrategia*. En esta categoría el estudiante no justifica o razona la respuesta, aunque como hemos comentado al principio de esta sección, es probable que hayan utilizado el coeficiente de correlación, como parece deducirse de la siguiente respuesta: *La relación que existe entre las dos variables es directa con intensidad fuerte.*

Estepa y Batanero (1996) pidieron a estudiantes que no había recibido instrucción sobre la correlación que estimasen el tipo de relación existente entre dos variables, dado el diagrama de dispersión, obteniendo un 47.5% de respuestas correctas. En nuestro caso la proporción de respuestas correctas es muy superior (74.1%), y la mayoría de estudiantes interpretó correctamente la correlación, a pesar que los datos se dan en forma de tabla, aunque los estudiantes habían recibido instrucción sobre la correlación. La justificación más utilizada se ha basado en el coeficiente de correlación, seguido de la covarianza y la variación conjunta. Es de destacar la confusión entre el sentido e intensidad de la correlación que manifiestan algunos alumnos.

Predicción de los Valores de la Variable Explicada a partir de la Variable Explicativa

El apartado 4 d) del problema 2 se refiere al uso de la recta de regresión para predecir, cuando se debe hacer una interpolación.

Los procedimientos de resolución utilizados en las respuestas a la pregunta sobre interpolación se exponen en la tabla 5. Han respondido 162 alumnos (83.93% del total). Aunque el valor obtenido de los cálculos es 31.625 bacterias, hemos considerado como respuesta correcta 31 ó 32 bacterias, ya que la respuesta a la pregunta formulada debe tener sentido en el contexto del problema, considerando erróneas las respuestas que ofrecen un valor decimal, lo que explica las pocas respuestas correctas, ya que los alumnos han prestado

menos atención al contexto y al resultado que ofrecen, que al modelo estadístico y los cálculos.

Tabla 5. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 2.d según procedimiento de resolución

Procedimiento	Tipo de respuesta			Número de alumnos que usa el procedimiento
	Correcta	Incorrecta	No responde	
Recta regresión y sobre x	5 (3.5)	137 (96.5)		142
Recta regresión x sobre y		8 (100)		8
Uso de proporcionalidad		10 (100)		10
No especifica estrategia		2 (100)		2
No responde			31 (100)	31
Total	5 (2.6)	157 (81.3)	31 (16.1)	193

Se puede observar otra vez la confusión entre las variables dependiente e independiente, en los alumnos que utilizan la recta de regresión de *x sobre y*. En otros casos se expresa una creencia en la existencia de proporcionalidad entre las variables, que se manifiesta por el uso de una regla de tres, como en la siguiente respuesta: *Si a 2 le corresponde 21 a 2.5 le corresponderá x.* Otro caso interesante es el alumno que observa que 2.5 es el punto medio del intervalo [2,3] y razona que la respuesta estará en el punto medio del intervalo correspondiente [21,33].

Predicción a partir de la Recta de Regresión de X sobre Y

Tal y como están planteadas las preguntas del problema 2, el objetivo de la cuestión e) era indagar si el alumno es consciente de la existencia de las dos rectas de regresión y si hace un uso adecuado de ellas, pues en Batanero, Godino y Estepa (1991, 2001) algunos estudiantes no discriminaban estas dos rectas En la tabla 6 mostramos los resultados.

Tabla 6. Frecuencia y (porcentaje) de soluciones correctas e incorrectas al problema 2.e según procedimiento de resolución

Procedimiento	Tipo de respuesta			Número de alumnos que usa el procedimiento
	Correcta	Incorrecta	No responde	
Recta regresión x sobre y	49 (50.5)	46 (47.4)		95
Recta regresión y sobre x		45 (100.0)		45
Uso de proporcionalidad		9 (100.0)		9
Otras estrategias		2 (100.0)		2
No especifica estrategia		2 (100)		2
No responde			40 (100)	40
Total	49 (25.4)	104 (53.9)	40 (20.7)	193

153 estudiantes (79.3% de la muestra), responden a la pregunta, la mayoría de los cuales (62.1%) utilizan la recta *x sobre y,* y en consecuencia han discriminado las dos rectas de regresión y la aplican de manera pertinente, salvo en errores de cálculo o a usar una expresión incorrecta de la recta de regresión.

Los que usan la recta de regresión *y sobre x* no son conscientes de la existencia de dos rectas diferentes ni de la utilidad de la recta de regresión de *x sobre y*. Los que emplean un procedimiento proporcional, tienen la creencia de que la recta de regresión es una aplicación lineal, hecho que, en general no es correcto, y suelen establecer una regla de tres. En

consecuencia, se aprecia un razonamiento proporcional con ciertas carencias. Dentro de éstos, hay una minoría que comparan el resultado obtenido en el apartado d) y, sin hacer ningún cálculo, responden que el tiempo que deberá transcurrir es de 2.5 horas, el mismo que el ofrecido en dicha cuestión. Esto es indicativo de una concepción determinista de la dependencia aleatoria (Estepa y Batanero, 1996) además de no discriminar las dos rectas de regresión.

5. SÍNTESIS DE RESULTADOS E IMPLICACIONES

En este estudio exploratorio hemos analizado los procedimientos de los estudiantes de la muestra cuando se enfrentan a la resolución de los problemas simples de regresión y correlación, destacando sus aciertos, errores y dificultades, cuyo análisis nos ha proporcionado una caracterización del conocimiento de los estudiantes sobre conceptos relacionados con la correlación y regresión analizados, al finalizar un curso de Estadística Descriptiva en los primeros cursos universitarios.

Casi las dos tercera partes de los estudiantes calcula correctamente el coeficiente de correlación, mientras que sólo la mitad llegan a una expresión correcta de la recta de regresión. A pesar de los errores de cálculo o uso de fórmulas incorrectas, muchos estudiantes reconocen que sus resultados no son correctos a la vista del diagrama de dispersión o de la magnitud del valor obtenido para el coeficiente de correlación. Tres de cada cuatro alumnos dan un juicio de asociación correcto, usando diversas estrategias correctas para tomar la decisión (coeficiente de correlación, covarianza, variación conjunta).

Los estudiantes tienen dificultades al tratar de relacionar conceptos matemáticos y estadísticos en la resolución del problema 1, por ejemplo, al interpretar los parámetros m y n en la ecuación de la recta, la ordenada en el origen o su centro de gravedad; estas dificultades disminuyen la capacidad de utilizar la recta de regresión con eficacia.

Un porcentaje elevado de estudiantes conoce el procedimiento matemático para efectuar las predicciones, pero ha habido errores de cálculo y sobre todo no se ha tenido en cuenta el contexto. Por otra parte, la confusión entre las variables dependiente e independiente da lugar, de manera sistemática y persistente a una serie de errores que dificulta la resolución adecuada de este tipo de problemas.

Las confusiones entre el sentido y la intensidad de la correlación, entre la variable dependiente e independiente (en varias cuestiones), o bien, la creencia de que la variación conjunta es proporcional son dificultades importantes a tener en cuenta en la planificación de la enseñanza.

A la vista de estos resultados, podemos concluir que este tema es complejo y una enseñanza deficiente puede dar lugar a concepciones erróneas y confusiones que obstaculicen una comprensión amplia del mismo, tan necesaria, tanto por su utilidad, como por ser requisito para seguir profundizando en el conocimiento estadístico. Estas dificultades nos alertan sobre la necesidad de reforzar la enseñanza de este tema, a través de situaciones didácticas apropiadas que permitan superar las dificultades observadas. Consideramos que sin una plena integración e interrelación de los diferentes conceptos y su interpretación dentro del contexto del problema, difícilmente los futuros graduados podrán llevar a cabo una práctica correcta de la estadística.

Dado que el estudio llevado a cabo es de tipo exploratorio, no podemos concluir con soluciones definitivas a los problemas detectados en el aprendizaje del tema, sino que estimamos que algunos de los resultados expuestos aquí precisan de nuevas investigaciones para quedar convenientemente clarificadas las causas de los errores, dificultades y concepciones erróneas que hemos detectado y, en consecuencia, poder proponer soluciones a

los problemas encontrados en la adquisición de conocimiento de los alumnos. Esto es especialmente importante puesto que estudios previos sugieren que los adultos tienen escasa capacidad intuitiva para estimar la correlación, salvo cuando ésta es muy fuerte y confirma sus teorías previas.

Agradecimientos: Este trabajo forma parte del Proyecto PB97-0851 (Ministerio de Ciencia y Tecnología, Madrid).

REFERENCIAS

Barbancho, A. G. (1973). *Estadística elemental moderna*. Barcelona: Ariel (Cuarta edición. Reimpresión de 1975).

Batanero, C., Estepa, A., Godino, J. D. y Green, D. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education,* 27(2), 151-169.

Batanero, C., Godino, J. D. y Estepa, A. (2001). La construcción del significado de la asociación mediante actividades de análisis de datos: Reflexiones sobre el papel del ordenador en la enseñanza de la Estadística. En J. R. Pascual (Eds.), *II Simposio de la Sociedad Española de Investigación en Educación Matemática (SEIEM),* (pp. 151-169). Pamplona: Universidad Pública de Navarra.

Cook, T. D. y Campbell, D. T. (1979). *Quasi-experimentation. Design & Analysis Issues for Field Settings*. Chicago: Rand M^{c}.Nally.

Cruise, J. R., Dudley, R. L. y Thayer, J. D. (1984*). A Resource Guide for Introductory Statistics*. Dubuque, IO: Kendall / Hunt.

Erlick, D. E. y Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology*, 73(1), 9-14.

Estepa, A. y Batanero, C. (1996). Judgments of correlation in scatter plots: student's intuitive strategies and preconceptions. *Hiroshima Journal of Mathematics Education*, 4, 25-41.

Estepa, A. y Sánchez-Cobo, F. T. (1998). Correlation and regression in secondary school text books. En L. Pereira-Mendoza, L. Seu, T. Wee y W.K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics* (vol. 2, pp. 671-676). Voorburgo (Holanda): International Statistical Institute.

Estepa, A. y Sánchez-Cobo, F. T. (2001). Empirical research on the understanding of association and implications for the training of researchers. En C. Batanero (Ed.), *Training Researchers in the Use of Statistics* (pp. 37-52). Granada: International Association for Statistical Education e International Statistical Institute.

Hildebrand, D. K., Lang, J. D. y Rosenthal, H. (1977) *Analysis of Nominal Data*. Londres: Sage University Paper.

Huberman, A. M. y Miles, M. B. (1994). Data management and analysis methods. En N. K. Denzin y Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 445-462). Londres: Sage.

Jennings, D. L., Amabile, T. M. y Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases,* (p. 211-230). Nueva York: Cambridge University Press.

Lane D. M., Anderson, C. A. y Kellam, K. L.(1985). Judging the relatedness of variables: The psychophisics of covariation detection. *Journal of Experimental Psychology. Perception and Performance*, 11(5), 640-649.

Miles, M. B. y Huberman, A. M. (1984). *Qualitative Data Analysis: A Sourcebook of New Methods.* Londres: Sage Publications.

Morris, E. J. (1997). *An investigation of students' conceptions and procedural skills in the statistical topic correlation*. Londres: The Open University, Centre for Information Technology in Education, Informe n. 230.

Moritz, J. (2002). Conflicting representations of statistical association. *Statistics Education Research Journal,* 1(1), 39-40.

Phillips, J. L. (1992). *How to Think about Statistics*. Nueva York: W.H. Freeman.

Sánchez-Cobo, F. T. (1999). *Significado de la correlación y regresión para los estudiantes universitarios*. Tesis doctoral. Departamento de Didáctica de la Matemática. 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.

Truran, J. M. (1997). Understanding of association and regression by first year economics students from two different countries as revealed in responses to the same examination questions. En J. B. Garfield y J. M. Truran (Eds.), *Research Papers on Stochastics Educations from 1997* (pp. 205-212). Universidad de Minnesota.

Vizmanos, J. R. y Anzola, M. (1988). *Matemáticas II. Opción C: Ciencias Sociales. Opción D: Humanística / Lingüística.* Madrid: SM.

ANTONIO ESTEPA
Universidad de Jaén
Departamento de Didáctica de las Ciencias
Campus Las Lagunillas s/n
23071 Jaén

IASE 2004 Research Round Table on Curricular Development in Statistics Education, Lund, Sweden, June 28 - July 3, 2004

The Round Table dates coordinate with those of the Tenth International Congress on Mathematical Education, which takes place in Copenhagen, Denmark 4-11 July 2004. Lena Zetterqvist (lena@maths.lth.se) and Ulla Holt will be local organisers. **Those interested** can contact Gail Burrill, Division of Science and Mathematics Education, College of Natural Science, Michigan State University, 116 North Kedzie, East Lansing MI 48824, USA, E-mail: (burrill@msu.edu).

IASE Activities at the 55th Session of the ISI, Sydney, Australia, April 5-12, 2005

Chris Wild is the IASE representative at the ISI Programme Co-ordinating Committee for ISI-55th Session, to be held in Sydney, Australia, April 5-12, 2005. As such he also is Chair of the IASE Programme Committee, which is in charge of preparing a list of Invited Paper Meetings to be organised by the IASE alone or in co-operation with other ISI Sections, Committees and sister societies. The committee will pay special attention to new topics that have been not discussed at the previous ISI Session. There is still time for you to propose a session theme for the IASE sessions for ISI55 in Sydney in 2005. Sessions that are of joint interest to IASE and another ISI section are also sought. Suggestions should normally include the name of the session organiser, a short description of the theme and an indicative list of possible speakers. Please email your proposals to Chris Wild at c.wild@auckland.ac.nz.

ICOTS-7, Working Cooperatively in Statistics Education, Brazil, 2006

We are also glad to announce that the IASE Executive accepted the proposal made by the Brazilian Statistical Association to hold ICOTS-7 in 2006 in Brazil. The proposal is also supported by the statistical associations in Argentina and Chile. Pedro Morettin <pam@ime.usp.br> is the Chair of the Local Organising Committee and Lisbeth Cordani <lisbeth@maua.br> is acting as a link between the IASE Executive and the local organisers. Scientific Committee IPC: Carmen Batanero (Chair), Susan Starkings (Chair Scientific Programme), John Harraway (Scientific Secretary), Allan Rossman and Beth Chance (Editors of Proceedings). More information from Carmen Batanero (batanero@ugr.es).