Experiencias Docentes

Didáctica con R. Menos cuentas y más pensamiento crítico

Didactics with R. Less calculations and more critical thought

Alejandro Galindo Alba

Revista de Investigación

Volumen VII, Número 1, pp. 053–074, ISSN 2174-0410
Recepción: 23 May'16; Aceptación: 1 Jun'16

1 de abril de 2017

Resumen

La estadística cada vez está tomando un papel más relevante en el desarrollo de la sociedad moderna. La crisis bursátil, las encuestas electorales, el manejo y clasificación de la información, la ciencia del dato o el Big Data son solo algunos ejemplos para entender la necesidad de tener una sólida cultura estadística para poder analizar nuestro entorno desde un punto de vista crítico y fundamentado.

Por todo ello, el objetivo de la siguiente comunicación es introducir una propuesta para la mejora de la didáctica de la estadística. Esta nos permitirá manejar un gran volumen de datos reales, evitar el uso de la calculadora, la visualización de gráficos y el análisis crítico de los resultados. Todo ello a través de un software libre, destinado hasta hoy a Estudios Superiores.

Palabras Clave: Cultura estadística, enseñanza y aprendizaje, investigación en educación estadística, R, software libre.

Abstract

Statistics is starting to play a more relevant role in the developement of modern societies. The stock market crash, election studies, management and classification of data, science of data or "Big Data" are only some examples that show how necessary is to have a solid statistical culture in order to analyse our environment with a critical point of view.

Due to of these reasons, the goal of the following paper is to make a proposal to improve Statistics didactics. That will allow us to handle a big volume of real data, make easier the interpretation of graphs, the critical analysis of results and reduce the usage of calculators. For that purpose we will use free software destined to superior studies nowadays.

Keywords: Statistics culture, teaching and learning, research in statistical education, R, free software.

1. Introducción

El objetivo de este trabajo es presentar una nueva herramienta para aquellos docentes que deseen introducir una opción de mejora en la didáctica de las matemáticas; particularmente en la estadística.

Conseguiremos:

- Evitar el tedioso trabajo con la calculadora.
- Trabajar con gran cantidad de datos.
- Trabajar con datos reales.
- Facilitar la visualización de los datos mediante gráficas.
- Fomentar la reflexión y el análisis crítico.
- Eliminar el gasto en software.

Para ello trabajaremos con R; un potente software gratuito que nos permitirá relacionar la utilidad de la estadística y la realidad del día a día.

A la vez, al eliminar los gastos en software también eliminamos una posible brecha económico-social. Este software no necesita ser instalado en ordenadores último modelo, ya que funciona en ordenadores menos actualizados realizando de igual manera sus funciones. Las rentas familiares no son un problema para que un alumno pueda instalar el software en casa.

2. El currículo de estadística

La estadística forma parte del currículo de matemáticas y es importante por su presencia en la sociedad (prensa, radio, política, etc.), en la enseñanza (obligatoria y no obligatoria) y en la investigación científica de diferentes ramas. No podemos imaginar un estudio experimental sin tener en cuenta la estadística. De este modo se posiciona como una herramienta fundamental en la sociedad de la información y se trata de dar una cultura estadística para todos como apunta Gal (2002).

A pesar de ser una materia sumamente importante y de la cual estamos rodeados a diario, la estadística suele ser una materia olvidada por los profesores en la educación secundaria, siendo relegada frecuentemente al final del temario. Los conceptos estadísticos quedan pobres, quizás por una escasa preparación del profesor en cuanto a la materia.

Por ejemplo, Holmes (1980) destacaba su importancia en las siguientes cuestiones:

- *La estadística es una parte de la educación general deseable para los futuros ciudadanos adultos, quienes precisan adquirir la capacidad de lectura e interpretación de tablas y gráficos estadísticos que con frecuencia aparecen en los medios informativos.*
- *Es útil para la vida posterior, ya que en muchas profesiones se precisan unos conocimientos básicos del tema.*

- *Su estudio ayuda al desarrollo personal, fomentando un razonamiento crítico, basado en la valoración de la evidencia objetiva.*
- *Ayuda a comprender los restantes temas de curriculum, tanto de la educación obligatoria como posterior, donde con frecuencia aparecen gráficos, resúmenes o conceptos estadísticos.*

Una visión más actual es la de Begg (1997), quien señala que la estadística es un buen camino para llegar a conseguir las capacidades de comunicación, procesamiento de la información, uso de nuevas tecnologías, trabajo cooperativo, que cada vez toman más peso en los nuevos currículos.

Las actitudes influyen decisivamente en el propio proceso de enseñanza y aprendizaje. Los profesores frecuentemente aíslan la estadística a un segundo plano, creando una actitud negativa hacia la misma. Esto lleva a los alumnos a no valorar la materia. Frecuentemente se la considera difícil, poco útil, se duda de su capacidad para aprenderla. Por lo tanto debemos tener en cuenta no solo la actitud de los estudiantes sino también las actitudes del profesorado. Quizás este bloqueo actitudinal viene heredado por su vínculo con las matemáticas.

Batanero (2002) por su parte, justifica la necesidad de la enseñanza de la estadística:

- *La estadística es una parte de la educación general deseable para los futuros ciudadanos adultos, quienes precisan adquirir la capacidad de lectura e interpretación de tablas y gráficos estadísticos que con frecuencia aparecen en los medios informativos.*
- *Es útil para la vida posterior, ya que en muchas profesiones se precisan unos conocimientos básicos del tema.*
- *Su estudio ayuda al desarrollo personal, fomentando un razonamiento crítico, basado en la valoración de la evidencia objetiva.*
- *Ayuda a comprender los restantes temas del currículo, tanto de la educación obligatoria como posterior, donde con frecuencia aparecen gráficos, resúmenes o conceptos estadísticos.*

3. La Estadística y R

Como señalaron Barriuso, Gómez, Haro y Parreño (2013), los alumnos adquieren las habilidades en el cálculo, pero no comprenden, en determinados casos, el sentido de lo que aprenden. Quizás les falta manipular los conceptos, es decir, verlos desde distintos ángulos para descubrir las relaciones que se dan entre ellos, para adquirir una comprensión más sólida de la estadística.

La tecnología y determinado software favorecen esta manipulación y la simulación de experimentos con mayor rapidez y fiabilidad. Si además este software es gratuito, facilitamos la accesibilidad al mismo tanto por parte de las administraciones como del alumnado, suponiendo un importante ahorro y eliminando a su vez posibles barreras económicas.

Por otro lado, acercar al alumno a este tipo de herramientas sin tener que esperar a llegar a estudios superiores les facilitará su manejo y comprensión en el futuro.

Para su demostración o inicio en su utilización, lo más interesante será aplicarlo a la estadística de las matemáticas de primero de Bachillerato; en ella se trata la estadística descriptiva bidimensional, la cual nos da bastante juego a la hora de plantear problemas, de ver la utilidad que tiene la estadística y de fomentar la visualización de los resultados mediante gráficas. Llevar esa materia a su aplicación en R nos permitiría trabajar con una gran variedad de datos, estudiar sus distribuciones unidimensionales, bidimensionales, distribuciones condicionadas, correlación de variables, regresión, estimación… y todo ello utilizando datos reales.

Desde un punto de vista más práctico, los alumnos podrían extraer datos reales de la página del Instituto Nacional de Estadística, volcarlos en R y utilizar las herramientas del programa para sacar conclusiones y demostrar la utilidad de su uso desde temprana edad.

3.1. Uso de las TIC en educación.

Las Tecnologías de la Información y la Comunicación (TICs) juegan en este principio de siglo un papel fundamental en la educación. Tratamos con generaciones que no han conocido el mundo sin internet y para los cuales un mundo sin tecnología es algo prácticamente impensable. Gracias a estas herramientas están acostumbrados a obtener información con facilidad fuera y dentro de la escuela, tienen una sorprendente capacidad de procesamiento paralelo, son altamente multimediales y al parecer, aprenden de manera distinta (OECD-CERI, 2006). Es por ello que los docentes también debemos atender esa llamada y plantear posibles mejoras en nuestro desempeño relacionando nuestros conocimientos con esta nueva ola generacional.

La escuela de hoy día, por tanto, se enfrenta a una necesidad de transformación; de una evolución desde la educación instrumental a otra que prepare a los estudiantes para desenvolverse en la sociedad del conocimiento. Las nuevas generaciones, que en determinados momentos, pasaran por el sistema educativo como si de una cadena de montaje se tratase, se preparan para puestos de trabajo que hoy todavía no existen y deben aprender a renovar permanentemente una importante parte de sus conocimientos y habilidades (21st Century Skills, 2002).

Una de las grandes fortalezas de las TIC en el aula de matemáticas es la posibilidad de utilizar software específico para acompañar la construcción del aprendizaje. Por otro lado, la utilización del software libre se apuntala cada vez más como la alternativa más idónea para el uso de las TIC en el ámbito educativo tal y como señalan Bracho y Maz (2012).

Como hemos visto, la irrupción y el uso generalizado de las TIC en los últimos años, está produciendo cambios de enorme importancia en las distintas áreas donde nos desenvolvemos y la educación no puede ser una excepción a pesar de que la escuela sea una de las instituciones más resistentes al cambio según España, Luque, Pacheco y Bracho (2008). Esto, unido a estudios realizados sobre este tema, concluyen que los estudiantes experimentan un aprendizaje significativo cuando usan adecuadamente las TIC en sus procesos de aprendizaje. Además, en el caso de las matemáticas, estos recursos ponen en manos de los profesores y estudiantes herramientas que contribuyen a desarrollar nuevas capacidades cognitivas, facilitan la comprensión de conceptos matemáticos, ayudan en la realización de cálculos complicados y facilitan el análisis en los procesos característicos de la resolución de problemas según Caravallo y Zulema (2009).

3.2. Ordenadores y enseñanza de la estadística.

Los ordenadores han jugado un papel fundamental en el desarrollo de la Estadística; tanto por facilitar el acceso a ella, como por proporcionar diferentes software sin los cuales hoy día sería inimaginable la realización de análisis de datos.

Autores y profesores como Shaughnessy, Garfield, & Greer (1996) ya tomaron consciencia de la importancia del uso de ordenadores en la enseñanza de la Estadística. Incluso el Instituto Internacional de Estadística ya realizó en Austría en 1970 y en Camberra en 1984 una Round Table Conference sobre el uso de computadoras en la docencia de la estadística.

Esta herramienta ha facilitado el trabajo con mayor número de datos reduciendo las horas de estudio gracias a su velocidad. Esto, permite a los alumnos investigar otros aspectos de los estudios estadísticos, como la recolección y planificación de la muestra, el diseño experimental y el análisis e interpretación de los resultados; es decir, tal como señala Batanero (2001), estas herramientas permiten que los estudiantes establezcan una relación con la estadística de la misma manera que lo podría hacer un estadístico profesional. Además, se acerca al alumno al manejo de la informática en general a través de procesadores de texto y hojas de cálculo.

La propia Batanero (2001) clasifica los siguientes tipos de software para la enseñanza de estadística:

- *Paquetes estadísticos profesionales, como por ejemplo: SPSS, STARTGRAPHICS, R, etc. Estos paquetes tienen una gran capacidad de tratamiento de datos y presentación gráfica.*
- *Software didáctico, como Fathom y Sampling Distributions.*
- *Software de uso general, como las hojas de cálculo EXCEL.*
- *Tutoriales.*
- *Software en Internet, material "on-line".*

3.3. Software libre.

Ofrecer la posibilidad de ser copiado, modificado, usado y distribuido de forma libre son las principales características del software libre que encuentran Sanchez y Toledo (2009). Este no ofrece los obstáculos del software comercial, ni imposiciones sobre sus licencias.

En cuanto a la relación con su uso en los centros educativos, podemos destacar que supone un importante ahorro en costes; dinero que en las escuelas no suele abundar y que puede destinarse a otros recursos. Pensemos que cuando en un centro educativo se monta un aula de informática se nos presentan dos opciones. Por un lado, la podemos equipar con software privado, que no podremos copiar ni modificar, por lo tanto los alumnos no podrán llevárselo a casa y el cual nos acarreará elevados costes. Por otro lado, podemos equipar la sala con software libre: Este podremos copiarlo libremente sin incurrir en el infringimiento de la ley, los alumnos lo podrán descargar en sus casas y estaremos exentos del costoso pago de las licencias.

Al eliminar los gastos en software también eliminamos una brecha económico-social. Este software no necesita ser instalado en ordenadores último modelo, sino que funcionan en ordenadores menos actualizados realizando de igual manera sus funciones. Las rentas familiares no son un problema para que un alumno se pueda llevar el software a casa.

3.4. R

R es el protagonista de esta propuesta didáctica. Es un software clónico del paquete S-Plus (no gratuito), diseñado especialmente para análisis estadísticos y gráficos por Ross Ihaka y Robert Gentleman.

En 1995 Martin Maechler convenció a sus creadores para que lo distribuyeran gratuitamente, pero hasta 1999 no llegaron las primeras versiones piloto que han ido evolucionando y añadiendo mejoras hasta la fecha.

Hasta hoy es uno de los paquetes estadísticos más utilizados en biomedicina, bioinformática, matemáticas financieras y series de datos bursátiles.

Existe mucha documentación relativa al uso del R en general. Puede consultarse las direcciones:

http://www.cran.r-project.org/other-docs.html

http://www.cran.r-project.org/manuals.html

En ellas se puede encontrar documentación en varios idiomas, entre ellos en español, sobre R y sus aplicaciones en el análisis estadístico.

Observación: al abrir R saldrá la línea de comandos, la cual comienza con el símbolo > . En los ejemplos de la comunicación se ha incluido este símbolo (el cual no debe ser tecleado si queremos ejecutar los ejemplos) para diferenciar los comandos de ejecución con las salidas que va devolviendo el programa.

Para la práctica en el aula se recomienda trabajar con la versión RStudio; facilitando así el manejo del programa a los estudiantes. En esta versión la pantalla de comandos, la consola y los gráficos aparecerán resumidos en una misma ventana tal y como se muestra en la figura 1:

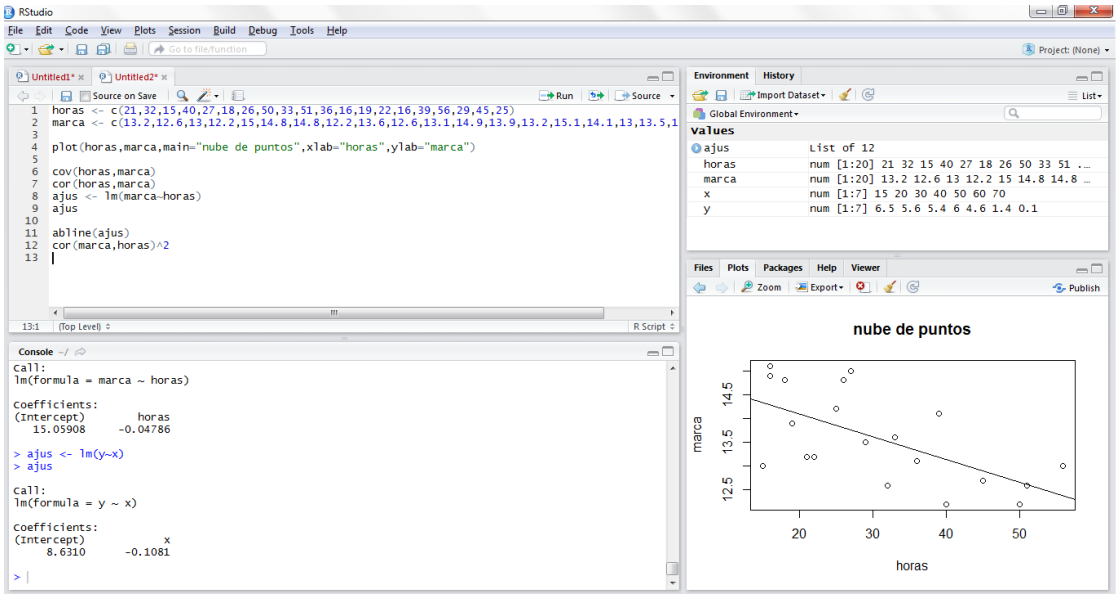


Figura 1. R-Studio.

4. Aplicación práctica

Siendo amplio el abanico de posibilidades para trabajar con R en la Educación Secundaria (obligatoria y post obligatoria), la investigación se enfocará en las matemáticas de 1º de Bachillerato.

4.1. Objetivos

Basándonos en el Real Decreto 1105/2014, de 26 de diciembre de 2015, por el que se establece el currículo básico de la Educación Secundaria Obligatoria definiremos los objetivos a conseguir de la siguiente manera:

- Encontrar valores representativos de un conjunto de datos utilizando medidas de centralización y dispersión.
- Representar e interpretar un conjunto de valores de dos variables mediante una nube de puntos.
- Distinguir si las variables de una distribución bidimensional tienen una relación de carácter aleatorio o funcional.
- Estimar el coeficiente de correlación lineal a partir de una nube de puntos.
- Analizar el grado de relación de dos variables de una distribución bidimensional conocido el coeficiente de correlación lineal.
- Determinar la recta que mejor se ajusta a la nube de puntos utilizando distintos procedimientos.
- Estimar un valor de una variable conocido un valor de la otra.

4.2. Competencias

Las competencias, o capacidades para activar y aplicar de forma integrada los contenidos del curso en el proyecto son cuantiosas. Las propuestas, para lograr la realización adecuada de actividades y la resolución eficaz de problemas complejos, que se han querido destacar son las siguientes: competencia en comunicación lingüística, a través del uso del vocabulario específico, de la comprensión de los enunciados o de la identificación de las diferentes formas en las que pueden venir representados los datos en un enunciado; la competencia matemática, distinguiendo los elementos que intervienen en la regresión lineal y correlación, la relación de las variables, la interpretación de las representaciones gráficas y la interpretación de los resultados; la competencia digital, ya que debe hacer un uso ético y crítico de las Tics; las competencias sociales y cívicas, aceptando a todos los componentes y opiniones del grupo; la competencia cultural, representando e interpretando la información con relación a su entorno; la competencia aprender a aprender, comprendiendo las lagunas en el aprendizaje a la vista de los problemas que se tengan para realizar estimaciones; y el sentido de la iniciativa y espíritu emprendedor, mostrando iniciativa al organizar las diferentes tareas o actividades a realizar, planificando su trabajo y mostrando iniciativa e interés por conocer y trabajar la curiosidad científica.

4.3. Recursos

Para el desarrollo de esta propuesta, el centro ha de contar con un aula equipada con ordenadores. Lo ideal es que cada alumno tenga su propio espacio y no sea necesario compartir los ordenadores para el trabajo individual.

El software se descargará de la página oficial y deberá estar instalado en cada equipo.

4.4. Metodología

Durante las sesiones se deberá crear un espacio de metodología activa que involucre tanto al docente como a los alumnos, donde estos puedan aprender a interpretar y visualizar los problemas de una manera autónoma y donde puedan ingeniar y experimentar técnicas propias para encontrar soluciones o rehacer las ya dadas. A su vez, se ha de buscar la colaboración entre ellos y el trabajo en grupo. Aprender a escuchar, discutir, argumentar y concretar objetivos comunes para un mismo logro, disfrutando y divirtiéndose, deberían ser parte de la metodología.

Para llevar esto a cabo, se trabajará con problemas que supongan un reto y un estímulo para el alumno; que le provoquen y le motiven. Problemas con un final poco frecuente o incluso con final abierto pueden provocar esa reacción que estamos buscando.

5. Fase de acción

5.1. Actividad 1

Esta actividad estará destinada a tomar contacto con la herramienta. Se entiende que los alumnos habrán recibido alguna clase previa de teoría y que aunque aún es pronto para dominar toda la terminología, sí le resultarán familiares términos como regresión, covarianza, correlación, etc.

El objetivo es que los alumnos vayan haciendo el ejercicio a la vez que el profesor, deteniéndonos en cada paso y analizando los resultados obtenidos.

Se midió el contenido de oxígeno, variable Y, a diversas profundidades, variable X, en el lago Wörthersee de Austria, obteniéndose los siguientes datos, en miligramos por litro:

X	*15*	*20*	*30*	*40*	*50*	*60*	*70*
Y	*6,5*	*5,6*	*5,4*	*6*	*4,6*	*1,4*	*0,1*

¿Podrías establecer una relación entre la profundidad del lago y el oxígeno observado a dicha profundidad?

Para desarrollar este ejemplo con R, lo primero que tendremos que hacer es incorporar los datos al sistema con los siguientes vectores:

```
> x <- c(15,20,30,40,50,60,70)

> y <- c(6.5,5.6,5.4,6,4.6,1.4,0.1)
```

Toda la laboriosidad del cálculo de las medias, varianzas, desviaciones típicas, etc. de cada variable no sería necesario en nuestro caso, ya que podemos obtener la recta de regresión, que aquí denominaremos *ajus*, al ejecutar el comando

```
>ajus<- lm(y~x)
```

A continuación, para ver los resultados y los coeficientes de regresión solo debemos ejecutar el objeto creado en el paso anterior.

```
>ajus

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept)          x

    8.6310      -0.1081
```

De esta manera obtenemos la recta de regresión ajustada de Y sobre X, que tiene por coeficientes los conseguidos en el paso anterior y que es

$$y = 8,6310 - 0,1081x$$

Con estos sencillos pasos ahorramos las tediosas operaciones con la calculadora y obtenemos resultados fiables para todos los alumnos. Hasta el momento, observando el coeficiente de la x: -0,1081, podríamos ver que las variables profundidad y oxígeno no están incorreladas y que la relación que se produce entre ellas sería de sentido inverso, es decir, cuando una de las variables crece, la otra decrece. Visto desde el punto de vista aplicativo, mientras mayor sea la profundidad a la que nos encontremos en el lago, menor será la presencia de oxígeno.

Para visualizar este comportamiento, podemos pedirle a R una representación gráfica de la nube de puntos:

```
>plot(x,y,xlab="profundidad",ylab="oxígeno")
```

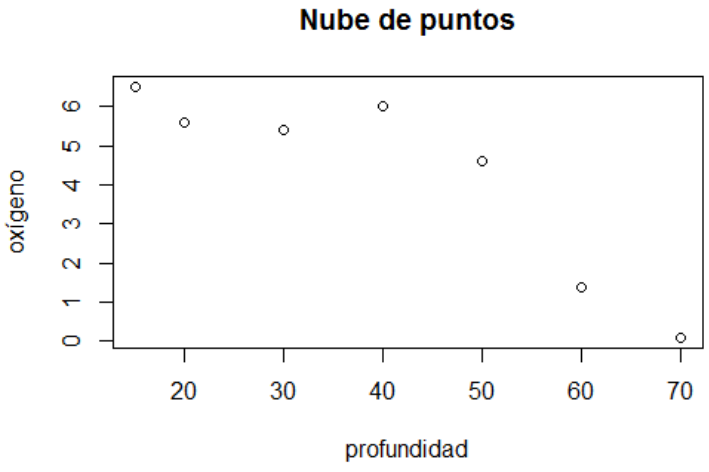


Figura 2. Nube de puntos (actividad 1).

Donde podríamos incluir la recta de ajuste para los datos ejecutando

>abline(ajus)

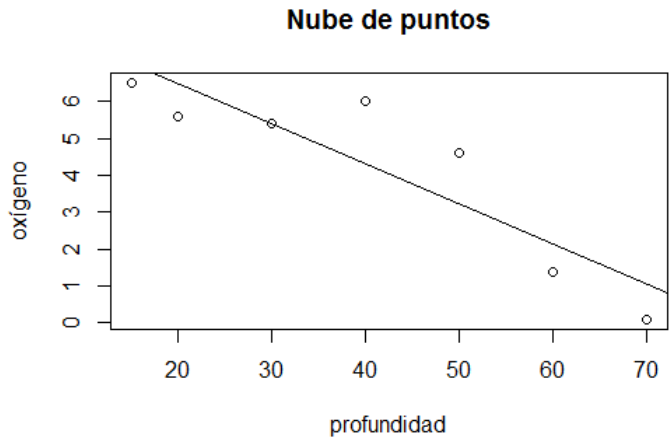


Figura 3. Recta de ajuste (actividad 1).

De esta manera, los alumnos podrán visualizar la relación entre las variables. Hecho que de otro modo sería no tan fácil de conseguir.

El inconveniente de la varianza como medida de relación es su dependencia de las unidades de medida de las variables. Por lo tanto, para trabajar con una medida adimensional la intensidad de relación lineal entre dos variables utilizaremos el Coeficiente de correlación lineal de Pearson.

$$r = \frac{Cov(x,y)}{S_x S_y}$$

Este coeficiente tiene el mismo signo que la covarianza y toma valores entre −1 y 1. Estos valores extremos reflejarían una relación lineal exacta entre las variables, lo que supondría que todos los puntos deben estar en una línea recta. Por lo tanto, cuanto más cercano a los extremos sea el valor del coeficiente mayor será la relación entre las variables, mientras que los valores cercanos a 0 indicarían una relación débil, por lo tanto no se podría explicar adecuadamente a la variable Y en función de X mediante la recta de mínimos cuadrados.

Para conocer su valor, ejecutamos

>cor(x,y)

[1] -0.8958494

El alumno, que ya sabía desde un principio que la relación entre las variables era de sentido inverso, puede observar ahora que dicha relación es bastante fuerte por la proximidad del coeficiente de correlación a -1.

Nos faltaría por conocer la fiabilidad del modelo, es decir, cómo de buenas son las predicciones, que calcularíamos mediante el modelo $y = 8,6310 - 0,1081x$.

El Coeficiente de Determinación es la herramienta que nos permite decidir si un ajuste es o no adecuado en sí mismo. Definido como

$$R^2 = \frac{Cov^2(x,y)}{S_x^2 S_y^2}$$

este coeficiente está comprendido entre 0 y 1, tratándose de un buen ajuste en aquellos casos donde R^2 esté cerca de 1, y de un ajuste deficiente en aquellos en los que sea cercano a 0.

Este coeficiente también lo podemos expresar como el cuadrado del coeficiente de correlación.

$$R^2 = \frac{Cov^2(x,y)}{S_x^2 S_y^2} = \left(\frac{Cov(x,y)}{S_x S_y} \right)^2 = (r)^2$$

Y de esta manera lo ejecutaremos:

>cor(x,y)^2

[1] 0.8025461

En nuestro caso se observa un buen ajuste del modelo. Esto quiere decir, que las estimaciones que hagamos a partir de él serán fiables.

Por ejemplo, para estimar la cantidad de oxígeno a 65 metros de profundidad, tan solo tendremos que sustituir 65 en la expresión de nuestro modelo $y = 8,6310 - 0,1081x$. En R se podría realizar la operación

> 8.6310-0.1081*65

[1] 1.6045

(Sería quizás conveniente que los alumnos repitan este ejercicio con pocos datos de manera manual. De este modo comenzarán a contrastar las virtudes del método y a valorarlo en su medida).

5.2. Actividad 2

En esta tarea hemos aumentado el número de datos y los alumnos trabajarán por parejas. Las funciones y el camino serán iguales que en el primer ejercicio, pero no así el resultado del estudio y sus posibles interpretaciones.

A medida que se avanza en la actividad se abrirá el debate entre las parejas sobre la decisión a tomar en cada paso del proceso, debiendo llegar a un consenso entre ambos.

En el centro de alto rendimiento "La Cartuja" de Sevilla se ha medido la marca que poseían 20 atletas en la prueba de 100 metros lisos y las horas semanales que, por término medio, le dedicaban a esta especialidad. Obteniéndose los siguientes resultados:

Horas	*21*	*32*	*15*	*40*	*27*	*18*	*26*	*50*	*33*	*51*
Marca	*13,2*	*12,6*	*13*	*12,2*	*15*	*14,8*	*14,8*	*12,2*	*13,6*	*12,6*
Horas	*36*	*16*	*19*	*22*	*16*	*39*	*56*	*29*	*45*	*25*
Marca	*13,1*	*14,9*	*13,9*	*13,2*	*15,1*	*14,1*	*13*	*13,5*	*12,7*	*14,2*

¿Sería eficiente dedicar recursos económicos para conseguir un modelo matemático a partir del cual estimar con fiabilidad la marca de un atleta a partir de las horas que entrena a la semana? Razonar la respuesta. En caso de respuesta afirmativa, calcular dicho modelo y estimar la marca de un atleta que entrena 54 horas semanales.

Como en cualquier ejercicio, lo primero que deberíamos hacer es introducir los datos en R. Para ello, en lugar de emplear los términos X e Y, emplearemos el nombre propio de cada variable. Esto nos ofrecerá versatilidad y comprobar que no todo está "cerrado".

> horas <- c(21,32,15,40,27,18,26,50,33,51,36,16, 19,22,16,39,56,29,45,25)

> marca <- c(13.2,12.6,13,12.2,15,14.8,14.8,12.2, 13.6,12.6,13.1,14.9,13.9,13.2,15.1,14.1,13,13.5, 12.7,14.2)

Y realizamos la representación gráfica de la nube de puntos

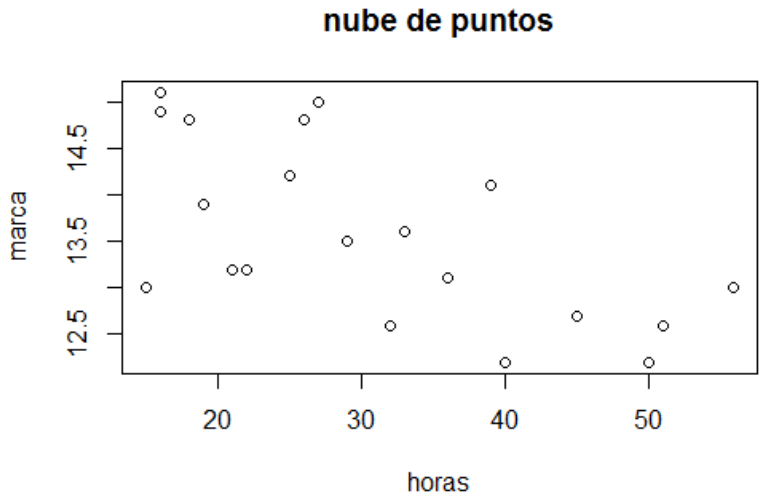


Figura 4. Nube de puntos (actividad 2).

Gracias a la importancia de la visualización gráfica, ya comenzamos a vislumbrar un posible comportamiento dependiente entre las variables. Para comprobar si las variables son o no incorreladas calcularemos la covarianza ejecutando

>cov(horas,marca)

[1] -7.650526

Comprobamos que la covarianza no es nula, luego las variables están correladas.

Mediante estos pasos se pretende que los estudiantes vayan estableciendo relaciones entre los conceptos teóricos y sus aplicaciones.

Hasta el momento habríamos podido comprobar que las variables no son incorreladas y que el sentido de su relación (que viene marcado por el signo de la covarianza) es de tipo inverso, es decir, cuando las horas aumentan las marcas mejoran.

En este punto los alumnos podrían pensar que ya es un buen momento para dedicar recursos económicos para conseguir un modelo a partir del cual predecir resultados. Y es el momento de incidir en la importancia del Coeficiente de Correlación, ya que es este el que nos marca la intensidad de la relación. Es decir, hasta el momento hemos encontrado que la relación existe, pero aún no sabemos cómo es de sólida.

Por lo tanto hay que calcular el coeficiente de correlación

>cor(horas,marca)

[1] -0.6304069

Al no ser un valor muy cercano a 1 o a -1, parece que la relación entre ambas variables, aunque existe, no es demasiado estrecha, por lo tanto no convendría desde un punto de vista económico emplear más recursos en el experimento. De cualquier modo se continuaría el ejercicio para mostrar otras curiosidades relevantes.

Si queremos obtener el modelo o recta de regresión de la marca sobre las horas entrenadas, ejecutamos

>ajus<- lm(marca~horas)

>ajus

Obteniendo la ordenada en el origen (15,05908) y la pendiente (-0,04785987)

Call:

lm(formula = marca ~ horas)

Coefficients:

(Intercept) horas

15.05908 -0.04786

La recta, por tanto, es

$$y = 15,05908 - 0,04785987x$$

Para añadirla a la nube de puntos ejecutamos la función abline

>abline(ajus)

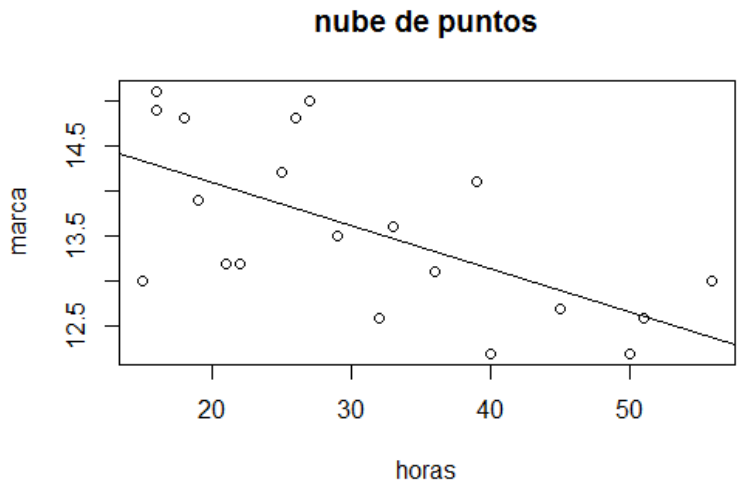


Figura 5. Recta de ajuste (actividad 2).

Es destacable la facilidad con la que hemos podido añadir la recta mediante la citada función.

Los alumnos podrían visualizar cómo la nube de puntos no parece demasiado concentrada alrededor de su recta de regresión. Esto debería comenzar a cobrar sentido, ya que vimos que las dos variables no presentaban una relación demasiado fuerte, lo que nos llevaría a pensar que las predicciones de las marcas que realicemos no serán muy fiables.

Una de las causas de esta falta de concentración de los datos puede deberse a que estamos estudiando una relación lineal entre las variables y estás puede que se ajusten mejor a otro tipo de modelo como el exponencial. Este es al tipo de reflexión que podrá llegar el alumno. Si lo piensa un poco, verá que no tiene demasiado sentido el modelo propuesto, ya que un atleta, por muchas horas que entrene nunca llegará a hacer una marca negativa.

La fiabilidad del modelo y por tanto de sus predicciones las calculamos con el coeficiente de determinación:

```
>cor(marca,horas)^2

[1] 0.3974129
```

Como era de esperar obtenemos una fiabilidad bastante baja, por lo que no merecerá la pena realizar predicciones.

Con esta actividad, los alumnos podrán experimentar con los conceptos aprendidos en las clases de teoría; aplicándolos a un caso real del que tendrán que extraer sus propias conclusiones.

Con el aumento del tamaño de la muestra los alumnos empezarán a comprender la utilidad de la herramienta y valorarán el tiempo que se ahorra en realizar las operaciones de cálculo a mano. Además, el trabajo en parejas fomentará el debate y la resolución de dudas.

5.3. Actividad 3

Esta tercera tarea será la primera que comiencen a trabajar por sí solos. Ya deberían conocer las herramientas que han de utilizar en cada momento, y lo más importante: a reflexionar sobre los resultados obtenidos.

Las calificaciones obtenidas, en dos asignaturas, por 17 alumnos de un centro escolar fueron las siguientes:

X	*3*	*4*	*6*	*7*	*5*	*8*	*7*	*3*	*5*	*4*	*8*	*5*	*5*	*8*	*8*	*8*	*5*
Y	*5*	*5*	*8*	*7*	*7*	*9*	*10*	*4*	*7*	*4*	*10*	*5*	*7*	*9*	*10*	*5*	*7*

¿Qué se puede decir acerca del coeficiente de correlación poblacional entre ambas variables?

5.4. Actividad 4

En esta tarea trabajaremos con un gran conjunto de datos perteneciente a R conocido como Fisher's iris data. Este recoge una muestra de 150 flores de tres especies diferentes de iris (iris setosa, versicolor y virginica) junto a sus variables longitud y anchura de sépalo (cm.) y longitud y anchura de pétalo (cm.).

Para acceder a los datos ejecutaremos directamente en R la función

>data (iris)

Una vez cargados podríamos ver los datos en la pantalla

>iris

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
…	…	…	…	…	…
146	6.7	3.0	5.2	2.3	virginica
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Figura 6. Datos (actividad 4).

La salida es un data.frame con 150 filas (observaciones) y 5 columnas (variables) llamadas Sepal.Length, Sepal.Width, Petal.Length, Petal.Width y Species.

A partir de aquí dividiremos la clase en grupos de 3-4 personas como máximo. La mitad de los grupos deberá estudiar la posible relación entre la longitud y anchura de los sépalos y la otra mitad de grupos la relación entre la longitud y anchura de los pétalos, reflexionando sobre los resultados obtenidos y extrayendo las conclusiones oportunas.

Para ejemplificarlo trabajaremos con el caso de los pétalos. Para ello, lo primero que tendremos que hacer es seleccionar los datos correspondientes a nuestro estudio.

> x <- iris[,3] # Longitud del pétalo

> y <- iris[,4] # Anchura del pétalo

Ejecutamos de manera esquemática para la demostración:

>plot(x,y,main="nube de puntos (pétalo)",xlab="longitud",ylab="anchura")

>abline(ajus)

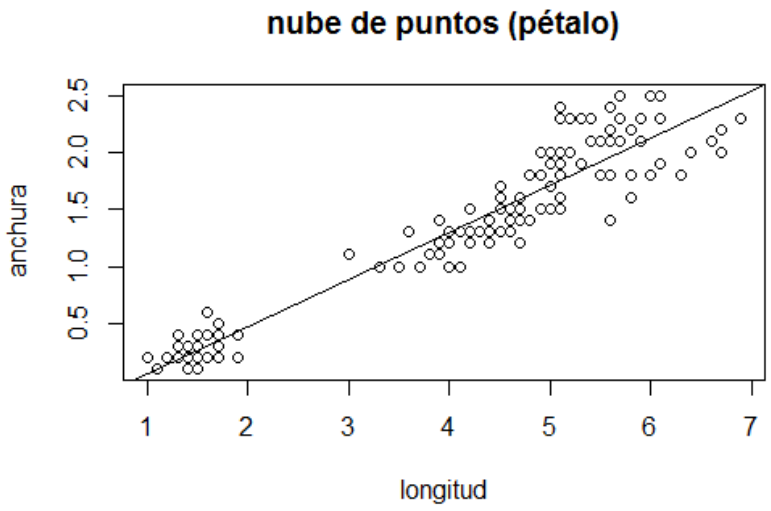


Figura 7. Recta de ajuste y nube de puntos (actividad 4).

```
>cov(x,y)
[1] 1.295609
>cor(x,y)
[1] 0.9628654
>ajus<- lm(y~x)
>ajus
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
   -0.3631       0.4158
>cor(x,y)^2
[1] 0.9271098
```

Con este ejercicio no solo hemos seguido desarrollando los contenidos, sino que podemos volver a relacionarlo con la realidad de nuestro entorno y con otras materias (inglés y biología).

La agrupación se ha llevado a cabo para desarrollar el trabajo activo y colaborativo de los alumnos. Derivando este en posibles debates entre los grupos en vista de los resultados.

5.5. Ejercicios de Autoevaluación.

A continuación se presentan cuatro actividades de autoevaluación para trabajarlas en pareja durante las dos últimas sesiones de la programación.

5.5.1. Autoevaluación 1

Esta primera tarea daría la oportunidad de relacionar la asignatura con otras materias como Geografía, a través de los Pirineos; o la Física y Química, a través de datos como presión atmosférica o punto de ebullición. Así mismo, en ella se trabajarán con pocos datos los primeros conceptos que trabajamos con R.

Se cree que existe una relación de tipo lineal entre el punto de ebullición del agua y la presión atmosférica del lugar en el que esta se pone a hervir. Para analizar esta hipótesis, se obtuvieron seis mediciones en Los Pirineos a seis alturas distintas en las que se observó una determinada presión atmosférica (en pulgadas de mercurio) X, anotándose la temperatura Y a la que comenzaba a hervir el agua (en grados Fahrenheit) en cada una de esas seis alturas. Los resultados obtenidos fueron los siguientes:

X	*20,68*	*22,42*	*23,91*	*23,99*	*25,09*	*29'10*
Y	*195,1*	*198,2*	*201,3*	*201,7*	*204*	*211,1*

a) Determinar la recta de regresión y analizar si es significativa.
b) Realizar el mismo estudio con los grados medidos en Centígrados.
c) Dibuja su nube de puntos.

5.5.2. Autoevaluación 2

Esta segunda actividad está diseñada para afianzar lo trabajado en la actividad anterior, pero con cuestiones que invitan más a la reflexión y a la demostración de la asimilación de los contenidos.

Se cree que el tamaño de los asentamientos prehistóricos puede servir para predecir el tamaño de la población del lugar donde se produjeron. Por ello se quiere determinar la recta de regresión basándose en datos actuales y, con ella, hacer estimaciones de tiempos pasados. Con este propósito se obtuvieron los siguientes datos de Tamaño de Asentamientos en hectáreas (X) y Número de habitantes (Y) de los pueblos actuales del área en estudio:

X	*0,7*	*1,1*	*1,2*	*1,3*	*1,7*	*2,0*	*2,4*	*3,1*	*3,2*	*3,4*
Y	*25*	*75*	*105*	*135*	*125*	*175*	*200*	*205*	*215*	*365*

X	*3,7*	*4,1*	*4,6*	*5,5*	*6,0*	*6,2*	*6,5*	*9,0*	*10,1*	*12,1*
Y	*305*	*255*	*505*	*275*	*195*	*635*	*655*	*315*	*735*	*855*

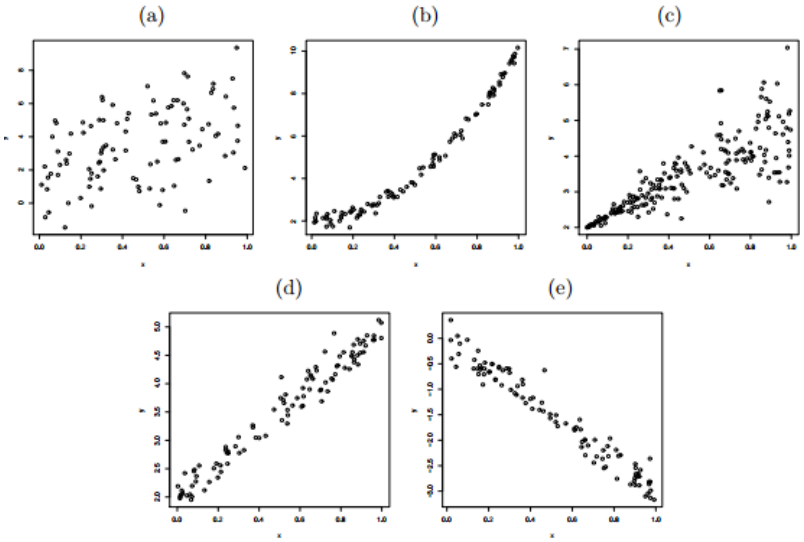
d) ¿Podríamos determinar si existe relación entre el tamaño de los asentamientos y el tamaño de la población?
e) ¿Podríamos intuir la relación sin realizar ningún cálculo numérico? ¿Cómo?
f) ¿Qué intensidad tendría dicha relación?
g) Determine la recta de regresión y analice si es significativa.

5.5.3. Autoevaluación 3

En esta actividad se pretende que los alumnos utilicen R como herramienta para la estimación e interpretación de modelos de regresión a través de la visualización de los gráficos.

Disponemos de un par de variables X e Y supuestamente relacionadas. A partir de una muestra de n individuos podemos visualizar la relación existente entre ambas. Responda a las siguientes cuestiones:

h) ¿Qué función de R hemos de utilizar para conseguir los gráficos de dispersión que se muestran abajo?
i) ¿Qué conclusiones podrías emitir sobre la relación de las variables en vista de las gráficas de cada ejemplo?
j) ¿Podrías plantear ejemplos reales en los cuales la relación entre las variables se ajusten a las gráficas?



5.5.4. Autoevaluación 4

En este último ejercicio de autoevaluación se plantea un reto a los alumnos, ya que deberán de trabajar con tres variables de dos en dos. Se pretende que demuestren que son capaces de interpretar la relación entre las variables, la fiabilidad de las mismas y que son capaces de comprarlas utilizando un lenguaje adecuado para la descripción de las situaciones.

El conjunto de datos trees, incluido en R, proporciona las medidas de la circunferencia en base (inches), altura (ft) y volumen (cubic ft) de 31 cerezos negros recién talados.

Realice un estudio acerca de la posible relación entre las variables presentando un informe con las conclusiones.

6. Conclusiones

Considero que la propuesta es arriesgada, comprometedora e innovadora; en donde quién sale ganando es el alumno por encima de todo. Arriesgada y comprometedora porque el profesorado que la imparta ha de tener la formación necesaria en programación en R para poder dirigir las clases con éxito; de otra manera sería inviable. También arriesgada porque hay que trasladar al grupo desde su aula al aula de informática, con el pertinente riesgo de distracción con los ordenadores, internet, fallos de conexión, etc.

Por otro lado, al empezar con las lecturas de bibliografía y búsqueda de información, no fue fácil encontrar casos que relacionaran la práctica de la estadística con R en educación secundaria; por eso me pareció una propuesta innovadora.

Con el manejo de R, no solo podremos dotar a los alumnos de una nueva herramienta con la que poder evolucionar en el aprendizaje de la estadística, sino que conseguiremos facilitar los cálculos de gran cantidad de datos para poder llegar de una manera más eficiente a los resultados; destinando de este modo el tiempo a trabajar reflexiones e interpretaciones de los resultados, a las posibles acciones a tomar en los estudios estadísticos y sin restarle importancia a la teoría.

Creo que es destacable el hecho de que los alumnos puedan conseguir fácilmente las gráficas, evitando así un problema tradicional en la enseñanza de la estadística. Esto nos permite el análisis visual de los datos, enriqueciendo los conceptos estudiados en las clases teóricas. Sin embargo, habría que ser cautelosos con esta herramienta para que no derive en un mal uso de la estadística.

Me parece un proyecto viable que requiere fundamentalmente de la implicación del docente. El alumno se dotará de unos recursos que le harán ver la estadística y todo lo relacionado con ella, desde un punto de vista crítico y analítico. Además, no lo olvidemos, utilizando un software libre, que elimina en todo momento una posible brecha social y económica, colocando a toda la comunidad educativa en igualdad de condiciones.

7. Propuestas de futuro

Una primera continuidad en el estudio la dirigiría hacia los primeros cursos de la etapa de secundaria. Por razones prácticas y demostrativas basé mi planteamiento en primero de Bachillerato, en el estudio descriptivo de dos variables, pero creo que sería interesante adelantar su aplicación a cursos anteriores.

De este modo también abriría otra línea de trabajo hacia su uso en la enseñanza de probabilidad, diseñando experimentos aleatorios, analizando el comportamiento de las variables, etc.

Sería interesante, en caso de que este proyecto se pudiera realizar, entrevistar a los alumnos que fueron formados en R y en una sólida cultura estadística, y estudiar los posibles beneficios que hayan encontrado con el paso del tiempo.

Referencias

[1] ARRIAZA, A. J., FERNÁNDEZ, F., LÓPEZ, M. A., MUÑOZ, M., PÉREZ, S., & SÁNCHEZ, A. *Estadística Básica con R y R-Commander*. Servicio de Publicaciones de la Universidad de Cádiz, Cádiz, 2008.

[2] BARRIUSO, J.M., GÓMEZ, V., HARO, M.J., & PARREÑO, F. *Introducción a la estadística con R*. Revista Suma. 72 (Marzo 2013), pp. 17−30., 2013.

[3] BATANERO, Carmen. *Estadística y didáctica de la matemática: Relaciones, problemas y aportaciones mutuas*. En C. Penalva, G. Torregrosa y J. Valls (Eds.), Aportaciones de la didáctica de la matemática a diferentes perfiles profesionales (pp. 95−120). Universidad de Alicante, Alicante, 2002.

[4] BATANERO, Carmen. *Didáctica de la Estadística*. Universidad de Granada, Granada, 2001.

[5] Begg, A. *Some emerging influences underpining assessment in statistics*. En I. Gal y J. Garfield (Eds.), The assessment challenge in statistics education. IOS Press, Amsterdam, 1997.

[6] BOE. *Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato*, 2015.

[7] CAMPBELL, S. *Flaws and Fallacies in Statistical Thinking*. Dover Publications, Nueva York, 2002.

[8] CARAVALLO, H. & ZULEMA, C. Z. *Herramientas para la enseñanza y el aprendizaje de las Matemáticas. Software libre*. En: Lapasta, L. (Ed.). II Jornadas de enseñanza e investigación educativas en el campo de las Ciencias Exactas y Naturales. Universidad Nacional de La Plata, 2009.

[9] ESPAÑA, F., LUQUE, C.M., PACHECO, M., & BRACHO, R. *Del lápiz al ratón. Guía práctica para la utilización de las nuevas tecnologías en la enseñanza*. Toro Mítico, Córdoba, 2008.

[10] ESTEPA, A. *Algunas notas sobre la didáctica de la estadística*. Junta de Andalucía, Jaén, 1993.

[11] FEBRERO, M., GALEANO, P., GONZÁLEZ, J. & PATEIRO, B. *Prácticas de Estadística con R*. Universidad de Santiago de Compostela, Compostela, 2012.

[12] FLORES, P. *Conceptos y creencias de los futuros profesores sobre las matemáticas, su enseñanza y aprendizaje*. Editorial COMARES, Peligros (Granada), 1998.

[13] GAL, I. *Adult's statistical literacy. Meanings, components, responsibilities*. International Statistical Review, 70(1), 1−25, 2002.

[14] GARCÍA, A. *Estadística básica con R*. UNED, Madrid, 2010.

[15] HOLMES, P. *Teaching Statistics* 11-16. Sloug: Foulsham Educational., 1980.

[16] IHAKA R. & GENTLEMAN R. (1996) *R: a language for data analysis and graphics*. Journal of Computational and Graphical Statistics 5: 299–314, 1996.

[17] OECD-CERI, (2006). *21st Century Learning: Research, Innovation and Policy*. Center for educational research and innovation. 2006. Disponible en: http://www.oecd.org/site/educeri21st/40554299.pdf [Fecha de acceso 28 de enero de 2016]

[18] PEÑA, D. *Fundamentos de Estadística*. Alianza Editorial, Madrid, 2008.

[19] RICO, Luís. *La educación matemática en la educación secundaria*. I.C.E. Universidad de Barcelona, Barcelona, 1997.

[20] ROJANO, T. *Incorporación de Entornos Tecnológicos de Aprendizaje a la Cultura Escolar: proyecto de innovación educativa en matemáticas y ciencias en escuelas secundarias públicas de México*. Revista Iberoamericana de Educación, 33, pp. 135-165., Méjico, 2003.

[21] RUÍZ, J. *Posibilidades de las TIC en el área de Matemáticas. En: Las TIC en la enseñanza y aprendizaje de las matemáticas*, pp. 11-19. MAD S.L. Alcalá de Guadaira, Sevilla, 2012.

[22] SÁNCHEZ, J. M., & TOLEDO, P. *Software libre y educación. En: El software libre en los contextos educativos*, pp. 11-26.MAD S.L., Alcalá de Guadaíra, Sevilla, 2009.

[23] SHAUGHNESSY, J. M., GARFIELD, J., & GREER, B. *Data handling. En A. Bishop et al. (Eds.), International handbook of mathematics education*, volumen 1, pp. 205-237. Dordrecht: Kluwer, A. P., 1996.

[24] SHUMWAY, R. H., & STOFFER. D. S. *Time Series Analysis and Its Applications: With R Examples*. Springer Science+Business Media, LLC., New York, 2011.

Sobre el autor:
Nombre: Alejandro Galindo Alba
Correo Electrónico: alegalalb@gmail.com
Institución: Departamento de Análisis Económico y Economía Política. Universidad de Sevilla, España.