DEL ANÁLISIS DE CORRESPONDENCIAS EN UNA INVESTIGACIÓN. APLICACIÓN E IMPLEMENTACIÓN CON R

Wilson Rodríguez, Claudia Daza y Felipe Fernández
Universidad Pedagógica Nacional (Colombia)
whrodriguezg@unal.edu.co, mdma_cldazag060@pedagogica.edu.co,
fjfernandez@pedagogica.edu.co

Por naturaleza los datos en muchos estudios son regularmente configurados a manera de tablas, que en asocio con los objetivos que se establezcan, determinan el tipo de análisis a seguir. Se presenta en este escrito una aplicación e implementación con R (Software libre y de uso frecuente en estadística), de un análisis de datos denominado en estadística Análisis de Correspondencias, precisado tal análisis por procesos como renglones y niveles como columnas de una tabla de datos obtenida en un experimento de enseñanza. El uso de R para la implementación del análisis se entiende como un aspecto complementario, para el mejor entendimiento mediante el uso de tecnologías de la informática. Se destaca que lo aquí contenido solo es un esbozo, en consecuencia los interesados en ahondar más en el tema deberán consultar algunas referencias que al final se registran.

PALABRAS CLAVE

Análisis de correspondencias, tecnología, procesos, niveles

INTRODUCCIÓN

Usualmente son conocidos métodos de análisis en los cuales los datos son del tipo cuantitativo, pero no es así cuando se cuenta con datos del tipo cualitativo, además, con regularidad las tecnologías informáticas cotidianas descuidan o poco implementan técnicas de análisis en el segundo tipo de datos. Hoy en día la naturaleza de la información obtenida en ciencias de la educación y otras ciencias, obligan a saber de alguna técnica que permita el manejo de datos cualitativos, y de manera particular cuando la disposición de los mismos esta en arreglos o configuraciones llamados tablas.

Aunque, ¿Cómo afrontar la información -o datos- contenida en una tabla? ¿Cómo interpretar una tabla de datos en su conjunto?, esto es, ¿Cómo leo la información en la tabla con respecto al objeto de estudio? y ¿Cómo se obtiene un resultado a partir de una tabla de datos, susceptible de ser verbalizado y comunicado por parte del investigador(es)?: son algunas de las preguntas que asisten a los investigadores cada vez que se enfrentan a tablas de datos.

Por las razones anteriores, esta ponencia pretende dar a conocer un método para el análisis de datos del tipo cualitativo en una investigación en educación, resaltando la aplicación en una situación problema y de manera más centrada en la implementación en el software R. Así a través de las sesiones: Marco de referencia, Desarrollo del tema y las Conclusiones del Análisis; se expondrán las distintas ideas para tratar de lograr

la pretensión inicial. Específicamente se realiza un esbozo de la metodología de análisis de datos cualitativos denominada en estadística 'Análisis de Correspondencias' (AC), basados totalmente en una investigación que adelantan los autores de este escrito.

En definitiva, se trata de comunicar y situar como objeto de debate la aplicación e implementación con R del análisis de correspondencias, de manera que se convierta la investigación-base en un espacio de reflexión y práctica no solo para investigadores, sino también para personas interesadas en la temática.

MARCO DE REFERENCIA

De entrada, diremos que el término *cultura estadística* o *statistical literacy*, hoy en día ha surgido entre los estadísticos y educadores en estadística, como forma de destacar que la estadística es considerada parte de la herencia necesaria para el ciudadano educado (Batanero, 2013). Esto es o implica cubrir de alguna manera la urgencia de formar personas estadísticamente alfabetizadas, y además capaces de pensar y razonar en la materia.

A condición de lo anterior, cada vez más el dialogo que existe entre el sujeto y los datos se debe hacer más claro, comenzando por el reconocimiento de la importancia de los datos, la configuración de estos y la adaptación de un modelo que permita describirlos, explorarlos y porque no llegar a un pronóstico.

Existen personas, grupos, asociaciones y otros que procuran poner de moda tal cultura. Baste, como muestra, las recomendaciones de la American Statistical Association (2012) que hicieron recientemente, a través del proyecto GAISE (Guidelines for assessment and instruction in statistics education) que como su nombre indica son unas directrices o guías para la instrucción y evaluación en educación estadística.

- Recomendación 1: Hacer hincapié en la alfabetización estadística y desarrollar el pensamiento estadístico.
- Recomendación 2: Utilizar datos reales.
- Recomendación 3: Lograr la comprensión conceptual, en lugar del mero conocimiento de los procedimientos.
- Recomendación 4: Fomentar el aprendizaje activo en el aula.
- Recomendación 5: Usar la tecnología para el desarrollo de conceptos y análisis de datos.
- Recomendación 6: Utilizar evaluaciones para mejorar y evaluar el progreso de los estudiantes.

Ahora, recogiendo lo más importante de lo antepuesto haremos mención al análisis de correspondencias, intentando no abandonar el encuadre en la cultura estadística. Díaz y Morales (2012) referencian el análisis de correspondencias en estadística, como una de las muchas técnicas del análisis multivariado con datos asociados a conjuntos de medidas, cuya naturaleza es cualitativa sobre un número de individuos u objetos. El conjunto de individuos junto con sus variables, pueden ser configurados en arreglos como matrices y/o tablas de dos o más entradas, por ejemplo, 6 áreas del

conocimiento registradas para 50 estudiantes en determinado curso; esta información se puede almacenar en una matriz de 50x6 o en una tabla cuyas filas o renglones sean los 50 estudiantes y por columnas las 6 áreas, se advierte que los títulos al margen de lo anterior no cuentan, aunque se deben usar en aras de la claridad.

También podemos decir que es una técnica de interdependencia, dicho de otra manera, es una técnica que busca el cómo y por qué se relacionan o asocian un conjunto de variables. De manera más concreta, este tipo de análisis está dirigido a tablas de contingencia e intenta conseguir la mejor representación simultánea de los dos conjuntos de datos contenidos allí (en las filas y en las columnas). Este análisis puede ser simple o bivariado si se confrontan dos variables y múltiple si son más de dos las cotejadas.

Para este trabajo, se adelantará el AC del tipo simple en una tabla que queda conformada por los procesos como renglones y los niveles como columnas. A manera de ejemplo, en la Tabla 1 se sugieren tres procesos en las filas y cinco niveles en las columnas. Estas características para nuestro caso serán medidas en un total de ocho grupos de estudiantes. Así, el objetivo será establecer la correspondencia entre procesos y niveles de las producciones de los estudiantes.

Si se compara este tipo de análisis, con otros que se pueden abordar para tablas de contingencia, se puede decir que el AC va más allá de solo hallar los grados de asociación entre variables (procesos y niveles) –que son expresados usualmente en números– para propender por una explicación del cómo y por qué de las asociaciones suscitadas.

Además, por considerar este tipo de análisis bien importante en el estudio adelantado y reconociendo que las imágenes en ocasiones ayudan en la claridad de las ideas, se muestra en las Figuras 1 y 2, tomadas de Pardo y Cabarcas (2001), las principales características del análisis referido. En la Figura 1: *X* corresponde a la tabla de doble entrada en donde hay *n* filas y *p* columnas. Así el término genérico x_{ij} corresponderá a la frecuencia absoluta que a la vez está en la fila *i*-ésima y *j*-ésima columna. De aquí se desprenden los perfiles o vectores fila y columna respectivamente que tienen sus coordenadas y se pueden representar en un espacio R^p y R^n a manera de nube de puntos. Por último (en la Figura 2), a dichas nubes se les halla la mejor proyección en un plano y se muestran los dos perfiles de manera conjunta, dejando ver las correspondencias que se presenten y la formación de agrupamientos entre filas y columnas. El rigor estadístico de este tipo de análisis se deja como anexo, en el trabajo base para esta comunicación.

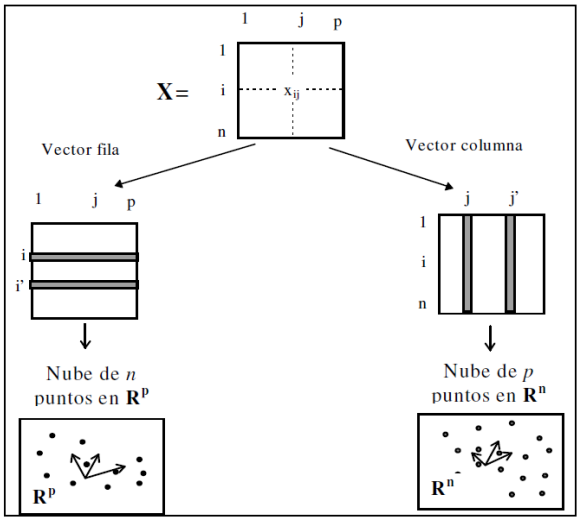

Figura 1. Perfiles y nubes de puntos

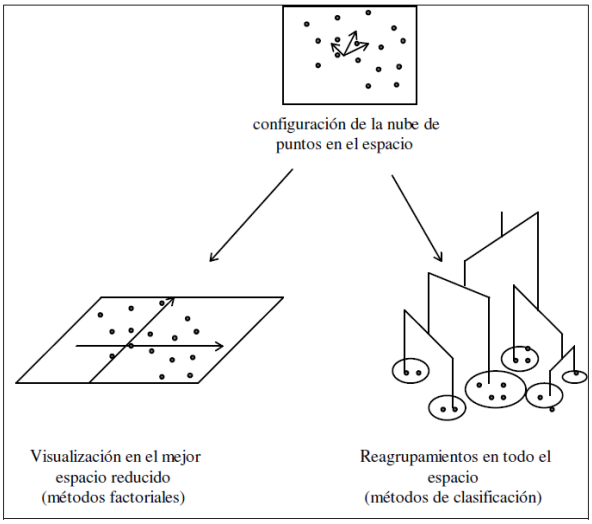

Figura 2. La mejor proyección

DESARROLLO DEL TEMA

Dominio del análisis

Se realiza un estudio denominado 'Exploración del razonamiento estadístico a nivel universitario bajo situaciones de muestreo' (Daza y Rodríguez, 2014), –tomado como base– en torno a la pregunta ¿Cómo es el razonamiento estadístico a nivel universitario bajo situaciones de muestreo? basado en la metodología de Experimentos de Enseñanza. La idea central es explorar el diálogo que establecen los estudiantes con los datos obtenidos por las situaciones aludidas. De manera más concreta, el estudio trata de examinar en estudiantes del nivel referido, el grado o alcance del sentido estadístico que pueden lograr a partir de datos muestrales.

El anterior trabajo se cimienta en una propuesta de tareas alrededor de una situación problema, dispuesta en cuatro partes o momentos para ser ejecutadas en dos sesiones de clase. Su desarrollo se hizo con estudiantes de tercer semestre del programa de

psicología, que se encuentran cursando la cátedra de Estadística inferencial en una universidad privada de Bogotá, periodo II-2013, donde se desempeña como docente uno de los investigadores. Se subraya que el número de estudiantes que participaron del experimento de enseñanza fue de 35, distribuidos en ocho grupos de trabajo (cinco grupos de cuatro estudiantes y tres de cinco).

Recogiendo lo más importante de lo anterior, se obtuvieron 136 respuestas, a 17 preguntas realizadas en cada grupo, que serán clasificadas en tres procesos y cinco niveles como se nuestra en la siguiente tabla.

		NIVELES				
		P	**U**	**M**	***R***	**AE**
PROCESOS	**I**	18	10	20	8	0
	II	15	23	5	4	1
	III	9	13	7	0	3

Tabla 1. Tabla de contingencia (Procesos y Niveles)

Las frecuencias absolutas allí contenidas son supuestas –hasta el momento de elaborar este escrito–, pues, aún no han sido consolidados tales resultados por parte de los investigadores, solo se usan las frecuencias evocadas debido a necesidad de continuar con el análisis. De modo que, por ejemplo, en la tabla la frecuencia 23 en el proceso II y en el nivel U, indica que hay 23 respuestas para el proceso II que fueron clasificadas en el nivel U de las 136 respuestas obtenidas.

A continuación se relacionan para una mayor comprensión y entendimiento, algo relacionado con los procesos y los niveles establecidos.

Procesos relacionados con el escenario del 'muestreo'

Los siguientes procesos son el resultado de algunas modificaciones hechas a las principales conceptualizaciones acerca del muestreo sugeridas por Watson (2005).

I. Proceso de entendimiento y aplicación de la terminología propia del muestreo en el contexto.

II. Proceso de posturas crítica frente al tamaño de la muestra y métodos de muestreo.

III. Proceso de consolidación de los hallazgos.

Niveles de razonamiento estadístico

El modelo de Biggs y Collis, (1982, citados en Jones *et al.*, 2005) ha sido consistentemente utilizado como la base de investigación para el estudio del pensamiento matemático y el razonamiento de los estudiantes en relación con diferentes objetos matemáticos como: los números, las operaciones con números, geometría y probabilidad. Este modelo está basado a partir de la estructura de resultados de aprendizaje observados, denominado taxonomía SOLO, la cual permite clasificar y evaluar el resultado de una tarea de aprendizaje, en función de su organización estructural.

Los niveles integrados en SOLO quedarían definidos del siguiente modo (Biggs y Collis, 1982, citados en Hernández, Martínez, Dafonseca y Rubio, 2005):

- *Nivel Prestructural (P)***.** Respuestas centradas en aspectos irrelevantes de la propuesta de trabajo, con contestaciones evasivas o tautológicas del enunciado.
- *Nivel Uniestructural (U).* Respuestas que contienen datos informativos obvios, los cuales han sido extraídos directamente del enunciado.
- *Nivel Multiestructural (M).* Respuestas que requieren utilización de dos o más informaciones del enunciado, las cuales siendo obtenidas directamente de éste, son analizadas separadamente, no de forma interrelacionada.
- *Nivel Relacional (R).* Respuestas extraídas tras el análisis de los datos del problema integrando la información en un todo comprensivo. Los resultados se organizan formando una estructura.
- *Nivel Abstracción expandida (AE).* Respuestas que manifiestan la utilización de un principio general y abstracto que puede ser inferido a partir del análisis sustantivo de los datos del problema y que es generalizable a otros contextos.

Implementación del análisis con R

Existen muchas tecnologías en informática que están disponibles para la estadística, software a modo de paquetes que facilitan la implementación de metodologías para el análisis de datos como SAS, SPSS, MINITAB, STATA y R. Entre los más usados, R (por recomendación de los autores) es en realidad uno de los más idóneos, pues, su manejo se considera relativamente fácil. Algunos aspectos ventajosos de R son:

- Software siempre disponible, es de uso libre.
- Un análisis soportado en R requiere varios pasos, así el interesado debe saber de ello, es decir, se requiere un usuario bien estructurado.
- Se pueden construir objetos (como funciones), por lo tanto propicia usuarios siempre analíticos y en procura de resolver problemas.

Entorno R

Veamos algunos apartes referentes a lo que es R y su relación con la estadística según Venables y Smith (1999). R es un colectivo integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.
- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R.

El término 'entorno' lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy

específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos. R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos. Como tal es muy dinámico y las diferentes versiones no siempre son totalmente compatibles con las anteriores. Algunos usuarios prefieren los cambios debido a los nuevos métodos y tecnología que los acompañan, a otros sin embargo les molesta ya que algún código anterior deja de funcionar. Aunque R puede entenderse como un lenguaje de programación, los programas escritos en R deben considerarse esencialmente efímeros.

Estadística y R

Se considera que muchas personas utilizan R como una técnica estadística. Venables y Smith (1999) prefieren describirlo como un entorno en el que se han implementado muchas técnicas estadísticas, tanto clásicas como modernas. Algunas están incluidas en el entorno base de R y otras se acompañan en forma de bibliotecas (packages). Junto con R se incluyen ocho bibliotecas (llamadas bibliotecas estándar), pero otras muchas están disponibles a través de Internet en CRAN (http://www.r-project.org). Como hemos indicado, muchas técnicas estadísticas, desde las clásicas hasta la última metodología, están disponibles en R, pero los usuarios necesitaran estar dispuestos a trabajar un poco para poder encontrarlas.

Existe una diferencia fundamental en la filosofía que subyace en R y la de otros sistemas estadísticos. En R, un análisis estadístico se realiza en una serie de pasos, con unos resultados intermedios que se van almacenando en objetos, para ser observados posteriormente, produciendo unas salidas mínimas. Sin embargo en paquetes como SAS o SPSS se obtendría de modo inmediato una salida copiosa para cualquier análisis, por ejemplo, una regresión, una prueba de hipótesis o un análisis discriminante.

Procesamiento de los datos con R

En primer lugar, la sintaxis (en letra cursiva) del análisis en R se irá consignando y a la vez se describirán las distintas ideas correspondientes a tal metodología. En segundo lugar el análisis se basa en los datos consignados en la Tabla 1.

Entonces el primer paso es cargar las librerías donde se encuentran las funciones que se usaran; esto lo realizamos con:

> library(MASS)
> library(rgl)
> library(ca)

Antes de realizar el análisis propiamente dicho, se deben introducir los datos, se calculan las frecuencias y se realizan los gráficos de los perfiles de la Tabla 1 junto con los respectivos títulos. Para ello se procede como sigue.

> t1=matrix(c(18,10,20,8,0,
+15,23,5,4,1,
+9,13,7,0,3),nrow=3,byrow=T)
> dimnames(t1)=list(PROCESOS=c("I","II","III")
+ ,NIVELES=c("P","U","M","R","AE"))

Si se quiere apreciar la tabla escribimos:
> t1

Se convierten las frecuencias absolutas en relativas de la Tabla 1, con sus totales marginales

*>round(addmargins(prop.table(t1)*100),2)*

Se hallan los perfiles de los procesos y se asignan a una tabla llamada t4

>round(addmargins(prop.table(t1,1)),4)
>t4=round(prop.table(t1,1),4)
>t4

Se grafican los perfiles de los procesos, pero antes, se abre un panel donde se consignaran dos gráficas más posteriormente.

>par(mfrow = c(1, 3))
>barplot(t(t4),beside = TRUE ,
+ col = c("lightblue", "mistyrose", "lightcyan",
+"lavender","cornsilk")
+,legend = rownames(t(t4)), ylim = c(0, 0.8))
> title(main = "Perfiles de los Procesos", font.main = 6)

Se hallan los perfiles de los niveles y se asignan a una tabla llamada t5

> round(addmargins(prop.table(t1,2)),4)
>t5=round(prop.table(t1,2),4)
>t5=as.table(t5)
>t5

Se grafican los perfiles de los niveles

> barplot(t5, beside = TRUE,
+col = c("lightblue", "mistyrose", "lightcyan"),
+legend = rownames(t5), ylim = c(0, 1))
> title(main = "Perfiles de los Niveles", font.main = 6)

Finalmente se grafican conjuntamente los dos tipos de perfiles (Procesos y Niveles), en un plano:

> plot(ca(t1), mass = c(TRUE, TRUE),main="Perfiles Conjuntos")

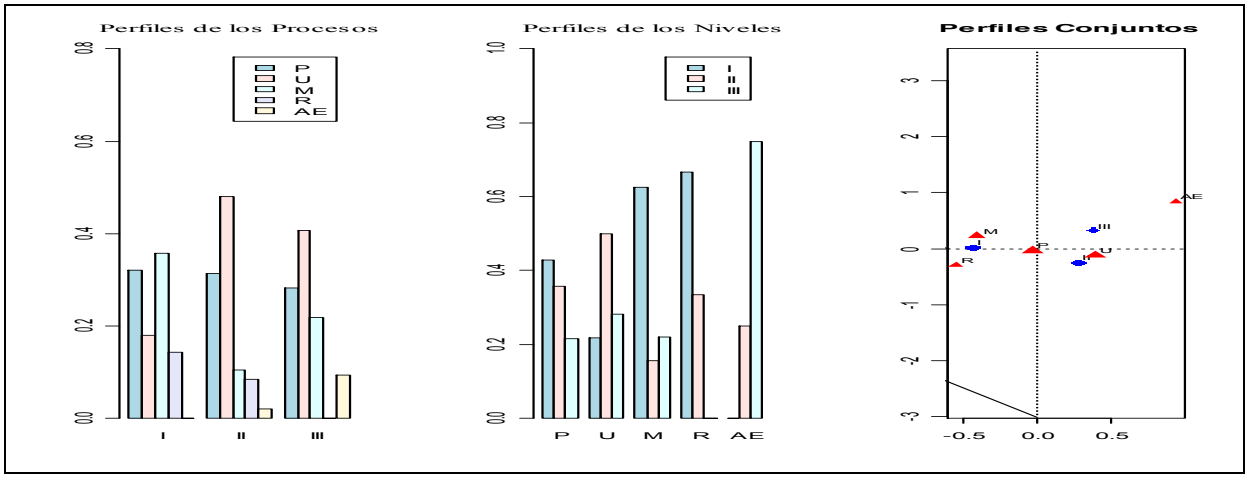

Figura 3. Panel de resultados, salida en R

CONCLUSIONES

La Figura 3 muestra el panel de las tres graficas que se programaron, así podemos obtener de allí algunas conclusiones.

En la primera gráfica

1. Se observa que en los tres procesos se tiene una representación del nivel P casi igual, 32%, 31% y 28% respectivamente, en tanto que los niveles AE y U son los más cambiantes para los mismos, por ejemplo AE con 0%, 2% y 10% respectivamente, además AE es el más bajo de los niveles para dos proceso. Así, podemos decir que, el nivel P no sirve para distinguir o caracterizar los procesos uno de otro, en tanto que el nivel AE o U sí lo harían.
2. El nivel U tiene la máxima representación con un 47% y 40% para los procesos II y III respectivamente.

En la segunda gráfica

3. En el nivel AE el proceso III tiene la máxima representación con un 75%. Parecería bueno, pero en realidad este nivel nunca es alcanzado en forma significativa en todos los procesos (ver Tabla 1).
4. En todos los niveles el proceso con más regularidad (uniformidad) es el II.

En la tercera gráfica

5. El nivel P aparentemente equidista de los procesos I, II, y III, ratificando de alguna manera lo dicho en la parte final de la conclusión 1.
6. En los procesos I y II el nivel máximo de compresión alcanzado por la mayoría de los grupos es el M y U respectivamente, en efecto esos niveles son los máximos en los correspondientes procesos (ver primera gráfica).
7. Los niveles *R* y AE aparecen con alto grado de oposición, evidenciando en las gráficas anteriores su comportamiento opuesto el uno del otro. Podemos interpretar aquí que tal oposición está dada por la generalidad que contiene el nivel AE y la particularidad del nivel *R* en cuanto a la comprensión alcanzada (ver definición de esos niveles).
8. El aislamiento de AE con respecto a todos los demás niveles y todos los procesos indica un aporte poco significativo al conjunto.

REFERENCIAS

American Statistical Association. (2012). *College report.* California: American Statistical Association.

Batanero, C. (2013). Sentido estadístico: Componentes y desarrollo. *I Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria* (pp. 2). Granada: Sociedad Española de Investigación en Educación Matemática.

Daza, C., y Rodríguez, W. (2014). Exploración del razonamiento estadístico a nivel universitario bajo situaciones de muestreo. Tesis de maestría. Bogotá: Universidad Pedagógica Nacional de Colombia.

Díaz, L. y Morales, M. (2012). *Análisis Estadístico de Datos Multivariados.* Bogotá: Universidad Nacional de Colombia.

Jones, G.A, Langrall, C., Mooney, E. y Thornton, C. (2005). Models of development in statistical. En D. Ben-Zvi y J. Garfield (Eds.), *The challenge of developing statistical literacy,*

reasoning and thinking (pp. 97-118). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Hernández, F., Martínez, P., Dafonseca, P. y Rubio, M. (2005). *Aprendizaje, competencias y rendimiento en educación superior.* Madrid: La Muralla.

Pardo, C. y Cabarcas, G. (2001). *Métodos estadísticos multivariados en investigación social.* Bogotá D.C: Universidad Nacional de Colombia.

Venables, W. y Smith, D. (1999). CRAN.R-project. http://www.r-project.org.

Watson, J.M. (2005). Developing reasoning about samples. En D. Ben-Zvi y J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.