EMP Educação Matemática Pesquisa
ISSN 1983-3156
A2
http://dx.doi.org/10.23925/1983-3156.2019v21i1p093-110

Iris de Fisher: sus posibilidades para un aprendizaje significativo de la clasificación y discriminación multivariantes

Iris de Fisher: suas possibilidades para uma aprendizagem significativa da classificação e discriminação multivariada

Iris de Fisher: its possibilities for a meaningful learning of the multivariate classification and discrimination

WAGNER, LAURA BEATRIZ[1]
TITIONIK, DIAMELA GISELLE[2]
DIESER, MARIA PAULA[3]
MARTÍN, MARÍA CRISTINA[4]
SCHLAPS, ÉRICA[5]
CAVERO, LORENA VERONICA[6]

Resumen

El conjunto de datos "Iris de Fisher" ha sido extensamente utilizado en la literatura estadística y en numerosos artículos sobre testeo y comparación de técnicas de discriminación y clasificación multivariadas. Sin embargo, los modelos creados a partir de estas técnicas, requieren el cumplimiento de ciertos supuestos que no son satisfechos por este conjunto de datos. El objetivo de este trabajo es presentar una propuesta para introducir los procedimientos del Análisis Lineal Discriminante y el Análisis de Agrupamientos (Clusters) utilizando estos datos clásicos, en un curso de análisis estadístico multivariado exploratorio, mediante el empleo del software R, con especial atención en el análisis de los supuestos necesarios, la estimación e interpretación de los modelos obtenidos, y la validación de resultados.

Palabras claves: *análisis lineal discriminante, análisis de clusters, software R.*

[1] Licenciada en Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa – lalywagner@gmail.com
[2] Licenciada en Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa – diamela2287@gmail.com
[3] Profesora en Matemática y Computación y Licenciada en Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa – pauladieser@exactas.unlpam.edu.ar
[4] Doctor of Philosophy in Science, Departamento de Matemática, Universidad Nacional del Sur – Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa – cristina.martin@uns.edu.ar; maritamartin@exactas.unlpam.edu.ar
[5] Licenciada en Matemática, Instituto de Ciencias Polares, Ambiente y Recursos Naturales, Universidad Nacional de Tierra del Fuego – ericaschlaps@gmail.com
[6] Profesora en Matemática, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa – cavero@exactas.unlpam.edu.ar

Resumo

O conjunto de dados "Iris de Fisher" tem sido amplamente utilizado na literatura estatística e em numerosos artigos sobre teste e comparação de técnicas de discriminação e classificação multivariada. No entanto, os modelos criados a partir destas técnicas, exigem o cumprimento de determinados supostos que não são satisfeitos por este conjunto de dados. O objetivo deste trabalho é apresentar uma proposta para introduzir os procedimentos de Análise Linear Discriminante e Análise de Agrupamentos usando estes dados clássicos, num curso de análise estatística multivariada exploratória, utilizando o software R, com especial atenção na análise dos supostos necessários, estimativa e interpretação dos modelos obtidos, e validação de resultados.

Palavras-chave: *análise linear discriminante, análise de agrupamentos, software R.*

Abstract

The "Fisher's Iris" data set has been extensively used in the statistical literature and in numerous articles on testing and comparing multivariate discrimination and classification techniques. The models created from these technique, require the fulfillment of certain assumptions, but these assumptions are not satisfied by this data set. The aim of this work is presenting a proposal to introduce the procedures of Linear Discriminant Analysis and Clusters Analysis using these classic data, in a course of exploratory multivariate analysis, using R software, with special focus on necessary assumptions analysis, the obtained models estimation and interpretation, and the results validation.

Keywords: *linear discriminant analysis, clusters analysis, software R.*

Introducción

El conjunto de datos “Iris de Fisher” ha sido extensamente utilizado en la literatura estadística y en numerosos artículos que se ocupan del testeo y comparación de diversas técnicas de discriminación y clasificación multivariadas (BEZDEK, KELLER, KRISHNAPURAM, KUNCHEVA y PAL, 1999; JOHNSON y WICHERN, 2007). Llama la atención, sin embargo, que los modelos creados a partir de estas técnicas, requieren el cumplimiento de ciertos supuestos que no son satisfechos por este grupo de datos clásico.

Varias son las razones pedagógicas por las que, a pesar de estas incongruencias formales de índole estadística, el conjunto sigue utilizándose en ambientes académicos para presentar dichas técnicas multivariadas. Estas razones pueden circunscribirse en torno a manifestaciones de diversos autores vinculadas a la necesidad de construir una cultura estadística. Según Gal (2002) es necesario promover el desarrollo de dos capacidades interrelacionadas: interpretar y evaluar críticamente la información estadística; y discutir o comunicar las opiniones respecto a tal información. Franklin, Kader, Mewborn,

Moreno, Peck, et al. (2005) indican que la enseñanza de la estadística debe ayudar a los estudiantes a aprender los elementos básicos del pensamiento estadístico: la importancia de los datos, la ubicuidad de la variabilidad, su cuantificación y explicación.

Wild y Pfannkuch (1999) afirman que el razonamiento estadístico es esencial para el aprendizaje e incluye cinco componentes fundamentales: reconocer la necesidad de los datos, la transnumeración, percibir la variabilidad, razonar con modelos estadísticos e integrar la estadística al contexto. Ante estos planteos, Batanero, Díaz, Contreras y Arteaga (2011) manifiestan que proponer a los estudiantes el desarrollo de actividades con datos reales, permite reemplazar la introducción de conceptos y técnicas descontextualizadas, o aplicadas a problemas tipo, difíciles de encontrar en la vida real, por una actividad integral motivadora que favorezca la construcción del conocimiento.

Este tipo de actividades requiere el uso de *software* específico. En los últimos años, el uso del lenguaje **R** se ha extendido globalmente entre investigadores de diversas áreas, consolidándose como uno de los programas estadísticos de referencia. **R** es libre, de código abierto, y multiplataforma. Dispone de una amplia variedad de técnicas y está en desarrollo constante, debido a numerosas contribuciones de la comunidad científica.

En este trabajo se presentan las principales etapas vinculadas con los procedimientos de aplicación de las técnicas de análisis discriminante y clasificación al conjunto de datos "Iris de Fisher", utilizando el *software* **R**. Estos procedimientos, seguidos en un curso de estadística multivariada exploratoria ofrecido como asignatura optativa a estudiantes de Licenciatura en Matemática, incluyen el análisis de los supuestos, la estimación e interpretación de los modelos obtenidos, y la validación de resultados.

El documento se estructura en cinco partes. En primer lugar se exponen algunos elementos teóricos que permiten fundamentar la propuesta didáctica que se describe brevemente en la segunda sección. A continuación se introducen las técnicas de discriminación y clasificación multivariadas estudiadas en la asignatura, seguidas de los resultados de la aplicación de las mismas sobre el conjunto de datos "Iris de Fisher". Para concluir, se realizan algunas reflexiones y consideraciones que justifiquen una posible reproducción, y eventual mejora, de esta propuesta en el futuro.

Fundamentación de la propuesta

La experiencia didáctica presentada en este trabajo pretende abogar por un aprendizaje significativo. A fin de promoverlo, el modelo de enseñanza situada se presenta

especialmente adecuado. En esta sección, presentaremos los elementos de las teorías de enseñanza y de aprendizaje mencionadas que permiten fundamentar la propuesta.

La teoría del aprendizaje significativo propuesta por David Ausubel se enmarca dentro de las denominadas teorías constructivistas. Según Ausubel (1976), el aprendizaje es significativo por definición, en tanto el estudiante relaciona de manera sustancial la nueva información (ideas culturalmente significativas) con sus experiencias y conocimientos previos (ideas de anclaje). Díaz Barriga (2003) afirma que, si se logra el aprendizaje significativo, es posible trascender la repetición memorística de contenidos inconexos y se logra dar sentido a lo aprendido, y entender su ámbito de aplicación y relevancia en situaciones académicas y cotidianas. En el mismo sentido, Pozo Municio y Pérez Echeverría (2009) resaltan que este tipo de aprendizaje facilita la generalización o transferencia en mayor medida que el aprendizaje repetitivo, e incrementa la capacidad de recuperar y usar esos conocimientos en nuevas situaciones.

No obstante, para que el aprendizaje significativo se produzca, es necesaria la disposición del alumno y la intervención del docente en esa dirección. En consecuencia, es sumamente importante la forma en que se plantean los materiales de estudio y las experiencias de aprendizaje. A fin de promover el aprendizaje significativo, resulta apropiado utilizar el modelo de enseñanza situada, derivado de las teorías de la cognición situada. Este enfoque instruccional, surge de la premisa de que el conocimiento es parte y producto de la actividad, el contexto y la cultura en que se desarrolla y utiliza (DÍAZ BARRIGA, 2003). Además, Pozo Municio y Pérez Echeverría (2009) sostienen que lograr una enseñanza superior eficaz, apoyada en enfoques constructivistas del aprendizaje, requiere orientar este último hacia la comprensión y fomentar un uso estratégico de los conocimientos adquiridos de manera que permitan resolver problemas o tareas nuevas, en lugar de limitarse a aplicar tales conocimientos en forma rutinaria.

En síntesis, la enseñanza situada destaca la importancia de la actividad y el contexto para el aprendizaje, entendiendo que aprender y hacer son acciones inseparables. Una estrategia de enseñanza que se enmarca dentro de este enfoque teórico es el aprendizaje basado en la solución de problemas auténticos. Ésta consiste en presentar situaciones reales o simulaciones auténticas vinculadas a la aplicación o ejercicio de un ámbito de conocimiento o ejercicio profesional, en las cuales el estudiante debe analizar la situación y elegir o construir una o varias alternativas viables de solución. Díaz Barriga (2003), sostiene que este tipo de experiencias favorecen una mayor retención y comprensión de

conceptos, aplicación e integración del conocimiento, motivación por el aprendizaje y desarrollo de habilidades de alto nivel.

Breve descripción de la experiencia de aprendizaje

El plan de estudios de la carrera Licenciatura en Matemática, incluida en la oferta académica de la Facultad de Ciencias Exactas y Naturales (FCEyN) de la Universidad Nacional de La Pampa (UNLPam), contempla una serie de asignaturas optativas y seminarios que los estudiantes deben realizar en los dos últimos años para alcanzar el grado correspondiente. Para quienes optan por la formación estadística, se ofrecen regularmente sendos cursos de análisis de datos multivariados desde un enfoque descriptivo y desde uno inferencial. Ambas asignaturas requieren el cursado previo de teoría de la probabilidad e inferencia estadística (obligatorias de tercer año).

El curso "Análisis de Datos Multivariados: Enfoque Descriptivo", de régimen cuatrimestral con una carga horaria semanal de 8 horas reloj, contempla el estudio de algunas técnicas exploratorias, análisis de *clusters*, análisis de componentes principales, discriminación y clasificación, análisis factorial, escalonamiento multivariado, y análisis de correspondencias.

La asignatura está organizada en clases teóricas y prácticas. En las primeras se discuten los fundamentos de las técnicas a estudiar acompañados de algunos ejemplos. Las segundas proponen la resolución de trabajos prácticos que incluyen ejercicios tipo y problemas reales. Algunos ejercicios permiten la resolución con lápiz y papel y están destinados a la familiarización del estudiante con la técnica a aplicar, mientras que otros requieren el uso de un *software* estadístico permitiendo el desarrollo de código para la obtención de la solución. Los problemas, en muchos casos atravesados por un enfoque histórico de la técnica bajo estudio, exigen al estudiante un análisis crítico de la situación y la elección de vías posibles de solución, aplicando los conceptos consolidados a partir de la realización de los ejercicios previos y fracciones de código ya desarrolladas. En estos intentos de solución, el estudiante puede encontrarse con nuevos problemas a tratar o la necesidad de desarrollar un código alternativo o adicional para resolver la situación propuesta.

Los problemas propuestos para el tratamiento de las técnicas de discriminación y clasificación involucran el análisis de un conjunto de datos clásico, conocido como "Iris de Fisher", utilizado a menudo en la literatura estadística sobre el tema. Obtener una

solución al problema, permite a los estudiantes comprobar que los modelos creados a partir de la aplicación de las técnicas estudiadas, requieren el cumplimiento de ciertos supuestos que no son satisfechos por este grupo de datos. Esto promueve la adopción de una posición crítica ante las soluciones ofrecidas a problemas estadísticos.

Técnicas de discriminación y clasificación

La característica principal del Análisis de Datos Multivariados (ADM) está en considerar un conjunto de *n* objetos sobre los que se observan los valores de *p* variables. El conjunto de objetos puede ser la totalidad o una muestra de un conjunto más grande donde las variables son cuantitativas o cualitativas. Para cada situación planteada, se busca estudiar este conjunto con diferentes propósitos.

En el problema de discriminación y clasificación se dispone de un conjunto amplio de elementos que provienen de dos o más poblaciones distintas. En cada elemento se han observado variables cuya distribución conjunta no necesariamente se conoce. Se desea diferenciar poblaciones o clasificar un nuevo elemento, con valores conocidos de sus variables, en poblaciones predefinidas. Dependiendo del tipo y del número, como así también de la distribución conjunta de las variables, existen diferentes enfoques para encarar este problema. Las técnicas pueden ser supervisadas o no supervisadas, según se conozca *a priori* o no, la categoría o clase a la que pertenece cada individuo de la muestra. Este trabajo se centra en el Análisis de *Clusters* y el Análisis Lineal Discriminante como técnicas no supervisada y supervisada, respectivamente.

Análisis de Clusters

Los numerosos procedimientos que pueden encontrarse bajo la denominación Análisis de *Clusters* (AClu) están basados en la idea básica de dividir los individuos de una población en grupos (*clusters* o conglomerados) de manera tal que exista homogeneidad interna entre ellos y aislamiento externo entre los grupos. Para ello se debe procurar, dependiendo del tipo de variables disponibles, una medida de disimilitud entre los individuos (euclideana, Manhattan, Mahalanobis son las más usuales, entre otras). Elegida ésta, se busca algún algoritmo que agrupe o clasifique los individuos con características similares (de OLIVEIRA BUSSAB, MIAZAKI y de ANDRADE, 1996; KAUFMAN y ROUSSEEUW, 1990).

Las técnicas de agrupamiento pueden clasificarse en dos: (a) Técnicas Jerárquicas, mediante las cuales los individuos son clasificados en grupos en diferentes etapas, de

modo jerárquico aglomerativo o partitivo (*Single*, *Complete*, y *Average Linkage*, Método del Centroide, entre otros), produciendo un árbol de clasificación denominado dendrograma; y (b) Técnicas no Jerárquicas, mediante las cuales los agrupamientos obtenidos en sucesivos pasos, producen una partición del conjunto original de individuos (Método de *k* - medias).

Análisis Lineal Discriminante

El Análisis Lineal Discriminante (ALD) propuesto por Fisher (1936), está basado en la Distribución Normal Multivariada de las variables consideradas y es óptimo bajo este supuesto (WELCH, 1939), siendo además necesaria la igualdad de matrices de covarianzas de los grupos considerados. La técnica consiste en dividir el espacio muestral en subespacios mediante hiperplanos que permiten separar lo mejor posible los grupos en estudio; para ello, a partir de la matriz de datos se construye una función (lineal) que será usada para discriminar y clasificar las unidades experimentales. A menudo los datos no son normales ni cumplen la condición de homocedasticidad, por lo que es necesario transformar las variables para aplicar el método.

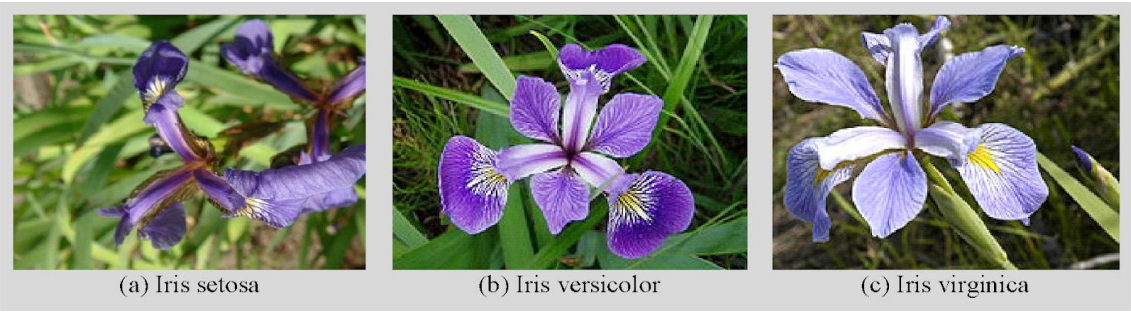
Aplicación del Análisis de Clusters y el Análisis Lineal Discriminante sobre los datos "Iris de Fisher"

A continuación se presentan los resultados obtenidos tras aplicar las técnicas antes descriptas (AClu y ALD) utilizando el lenguaje **R**, en su versión 3.2.2, sobre el conjunto de datos "Iris de Fisher". Dado que en la literatura circulan al menos dos versiones del conjunto de datos (BEZDEK, KELLER, KRISHNAPURAM, KUNCHEVA y PAL, 1999), cabe aclarar que en este trabajo se utilizan los datos del conjunto iris disponible en el paquete básico de **R** y publicados en Johnson y Wichern (2007).

El conjunto de datos

Los datos fueron recolectados, en 1935, por el botánico Edgar Anderson, con el objetivo de cuantificar la variación geográfica de estas flores en la Península de Gaspé, en Canadá (ANDERSON, 1935). Posteriormente, el biólogo y estadístico Ronald Fisher, utilizó estos datos, para desarrollar un modelo lineal discriminante que distinga las especies entre sí (FISHER, 1936). El conjunto de datos original consta de 50 muestras de cada una de las tres especies de iris (setosa, versicolor y virginica) sobre las que se midieron la longitud y el ancho de los sépalos y pétalos, en centímetros (Figura 1).

Figura 1 – Especies de Iris



Fuente: Wikimedia Commons.

Análisis de supuestos de normalidad y homocedasticidad

Primero se analiza la normalidad para cada una de las variables por separado y luego por pares, triplas y por último para las cuatro variables juntas para cada una de las especies, siguiendo el procedimiento delineado en Johnson y Wichern (2007).

Para analizar la normalidad univariada se puede realizar un histograma que brinda una idea visual de la distribución y se formaliza inferencialmente según el *Test de Shapiro-Wilk*. Este test utiliza la función **shapiro.test** de **R**. La normalidad bivariada también puede ser analizada gráficamente, mediante la función **scatterplot** de **R**, observando si los datos se circunscriben en una elipse. El análisis inferencial se realiza, nuevamente, por medio del *Test de Shapiro-Wilk*. En este caso, al igual que en los que se consideran la combinación de más de dos variables, se usa la función **mshapiro.test** de **R**.

Los resultados del *Test de Shapiro-Wilk* se muestran en la Tabla 1, indicándose los *p* - valores para cada posible combinación de las 4-variables y para cada especie en estudio. Se puede concluir que para un nivel de significancia del 10%, las variables Sepal.Length, Sepal.Width tienen una distribución normal para cualquiera de las tres especies, mientras que la variable Petal.Length es normal sólo para las especies versicolor y virginica y, la variable Petal.Width no está distribuida normalmente para ninguna de las especies. La normalidad para dos o más variables nunca se verifica en forma global.

Tabla 1: *p* - valores para cada una de las 15 posibles combinaciones de las 4 - variables y para cada especie de iris en estudio.

Combinaciones de variables	*p* – valor		
	Setosa	Versicolor	Virginica
Sepal.Length	0,4595	0,4647	0,2583
Sepal.Width	0,2715	0,3380	0,1809
Petal.Length	0,0548	0,1585	0,1098
Petal.Width	8,66e-07	0,0273	0,0869

Sepal.Length Sepal.Width	0,3017	0,3937	0,0821
Sepal.Length Petal.Length	0,4633	0,4438	0,7673
Sepal.Length Petal.Width	0,0011	0,3503	0,4168
Sepal.Width Petal.Length	0,4024	0,6515	0,0021
Sepal.Width Petal.Width	0,0001	0,0049	0,1349
Petal.Length Petal.Width	2,11e-05	0,3206	0,3154
Sepal.Length Sepal.Width Petal.Length	0,1412	0,3557	0,0029
Sepal.Length Sepal.Width Petal.Width	0,0011	0,0007	0,0244
Sepal.Length Petal.Length Petal.Width	0,0012	0,3710	0,9070

Fuente: Resultados obtenidos mediante R.

Para verificar la homocedasticidad de las variables con respecto a cada especie se utiliza la prueba de Bartlett para el caso univariado, invocando la función **bartlett.test** de **R**. Para el caso multivariado la prueba de M de Box es el recurso utilizado, a través de la función **BoxMTest** en **R**. Las 15-combinaciones de posibles *tests* arrojaron siempre el valor $p = 0$, lo cual muestra la falta de homocedasticidad entre estos datos.

Se incluye en la Figura 2 el código que permite llevar adelante el análisis exploratorio del conjunto "Iris de Fisher" disponible como paquete de datos de **R**. Las sentencias corresponden al estudio de la variable Sepal.Length para la población setosa. El lector puede reproducir el *script* para otra combinación de variables y población.

Figura 2 – Script de R para desarrollar el análisis exploratorio de "Iris de Fisher"

```
data(iris,package="datasets")
subdatos1=subset(iris,iris[,5]=="setosa")
media=mean(subdatos1[,1])
desvio=sd(subdatos1[,1])
m=min(subdatos1[,1])
n=max(subdatos1[,1])
hist(subdatos1[,1],main="Histograma",xlab="Sepal.Length",ylab="frecuen
     cia",plot=TRUE)
abline(v=c(media-2*desvio,media-
     desvio,media,media+desvio,media+2*desvio),lty="dashed",col="red")
shapiro.test(subdatos1[,1])
```

Fuente: Código propio.

El código que se muestra en la Figura 3 permite verificar la normalidad y la homocedasticidad de las variables. En la primera sentencia se analiza gráficamente la normalidad bivariada entre las variables Sepal.Length y Sepal.Width de la especie setosa. Luego, a través del *Test de Shapiro-Wilk*, se verifica la normalidad de las 4-variables para esta misma especie. En el cuarto comando se ejecuta la prueba de Bartlett para verificar

la homocedasticidad de la variable Sepal.Length con respecto a cada especie. La última línea permite realizar el test M de Box para todas las especies incluyendo las 4-variables. El lector puede ejecutar el *script* para otra combinación de variables y especie.

Figura 3 – Script de R para analizar la normalidad y homocedasticidad del conjunto "Iris de Fisher"

```
scatterplot(subdatos1[,1],subdatos1[,2],main="setosa",xlab="Sepal.Leng
    th",ylab="Sepal.Width",ellipse=TRUE,levels=c(.95),pch=20,boxplots=
    FALSE,smoother=FALSE,reg.line=ALSE,xlim=c(4,6),ylim=c(2,5))
library(mvnormtest)
mshapiro.test(t(subdatos1[,c(1:4)]))
bartlett.test(iris[,1],iris[,5])
BoxMTest(iris[,c(1,2,3,4)],iris[,5],alpha=0.05)
```

Fuente: Código propio.

Resultados del Análisis de Clusters

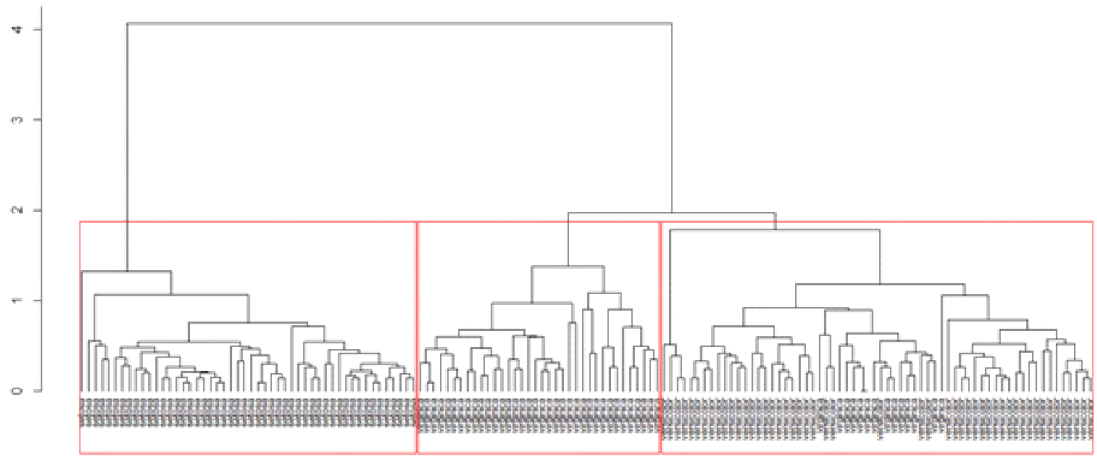
No existiendo un criterio de preferencia para decidir por una medida de disimilitud, se opta por trabajar con la distancia euclideana, concentrando la atención en tres métodos jerárquicos aglomerativos (*Single*, *Complete* y *Average Linkage*) y el método *k*–medias, como caso no jerárquico, partiendo de diferentes configuraciones iniciales.

Las técnicas jerárquicas seleccionadas agrupan los individuos siguiendo diferentes criterios. Mientras *Single Linkage* forma los grupos a partir de los miembros más cercanos, *Complete Linkage* los compone considerando aquellos más alejados. En cambio, *Average Linkage*, fusiona los grupos a partir de la distancia media entre pares de miembros de los grupos respectivos (JOHNSON y WICHERN, 2007). Para validar en qué medida, la estructura obtenida por cada procedimiento representa las similitudes o diferencias entre los objetos de estudio se utiliza el coeficiente de correlación cofenético (SOKAL y ROHLF, 1962). Éste mide la correlación entre las distancias iniciales, tomadas a partir de los datos originales, y las distancias finales con las cuales los individuos se han unido durante el desarrollo del método. Valores altos del coeficiente cofenético indican que durante el proceso no ha ocurrido una gran perturbación en la estructura original de los datos.

Cargados los datos **iris** y el paquete **cluster** de **R**, se utiliza la función **hclust** para obtener los agrupamientos resultantes de la aplicación de métodos jerárquicos. La función requiere al menos dos parámetros: una matriz de distancias entre observaciones y el método jerárquico a utilizar. Entre otros resultados, la función **hclust** proporciona el dendrograma mediante la función **plot**, mientras que el coeficiente de correlación cofenética se obtiene a partir de la función **cophenetic** del paquete **stats**.

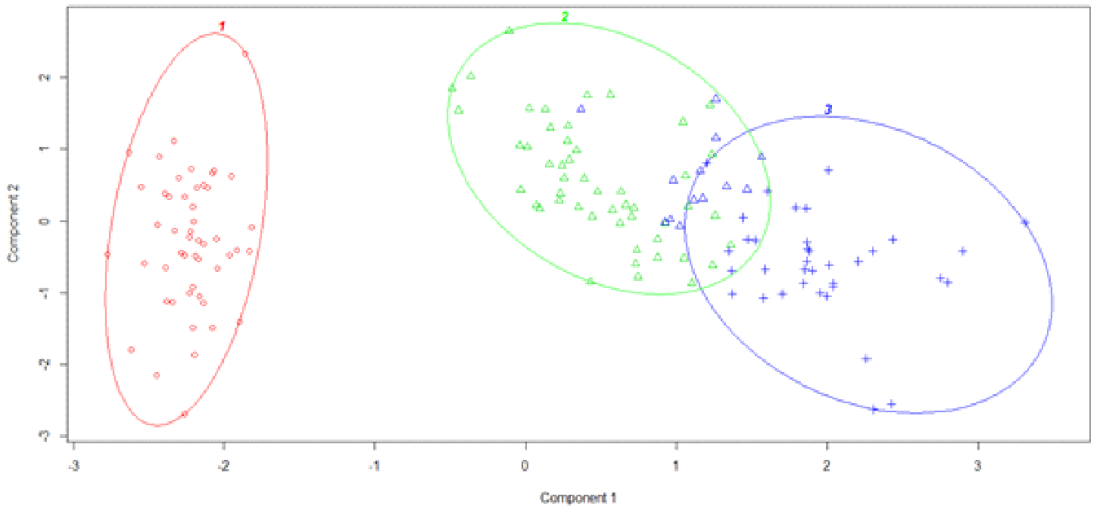
Las estructuras resultantes de la aplicación de los tres métodos poseen diferentes grados de perturbación respecto de la original. Los métodos *Single* y *Average Linkage* originan estructuras similares a la inicial (r = 0.8638787 y r = 0.8769561, respectivamente), permitiendo identificar dos grupos a partir del análisis de los dendrogramas resultantes. Por su parte, *Complete Linkage* evidencia una perturbación mayor (r = 0.7269857), pero generando tres *clusters*. Estas conclusiones se obtienen a partir de la determinación del paso en el que se produce la mayor variación en las distancias de unificación de *clusters*. Dado que *Average Linkage* arroja el valor más alto del coeficiente de correlación cofenética, se opta por los resultados obtenidos a partir de este método. El conocimiento *a priori* de la existencia de muestras provenientes de tres poblaciones, permite obtener los agrupamientos resultantes en el dendrograma de la Figura 4. Los resultados de esta partición también pueden observarse en el *clustplot* de la Figura 5. Las elipses encierran las observaciones agrupadas en un mismo *cluster*. Los símbolos usados para las observaciones identifican el grupo asignado por el método (○: grupo 1; △: grupo 2; +: grupo 3) y los colores indican la clasificación original (rojo: setosa; verde: versicolor; azul: virginica). Este gráfico representa las unidades muestrales relativas a las dos primeras componentes principales (explican 95.81% de la variabilidad total), y los grupos resultantes como elipses (PISON, STRUYF y ROUSSEEUW, 1999). Se observa que los ejemplares setosa se agrupan de manera perfecta, evidenciándose confusión entre las especies virginica y versicolor.

Figura 4 – Dendrograma de los 150 ejemplares de iris utilizando *Average Linkage* a partir de la matriz de distancias euclideanas (r = 0.8769561).



Fuente: Elaboración propia mediante R.

Figura 5 – *Clustplot* de los 150 ejemplares de iris usando *Average Linkage* a partir de la matriz de



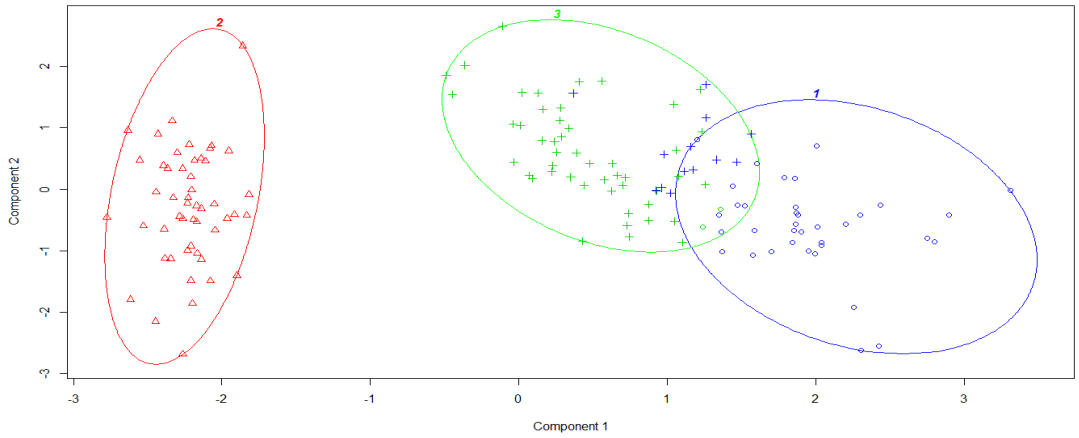
Fuente: Elaboración propia mediante R.

Por otro lado, el método *k* - medias, a diferencia de los métodos jerárquicos, no requiere del cálculo de una matriz de distancias inicial, aunque sí es necesario determinar el número de grupos a conformar, una agrupación inicial y el algoritmo de actualización de grupos. Forgy (1965) y McQueen (1967) proponen elegir aleatoriamente *k* centros iniciales del conjunto de datos originales y asignar cada individuo al grupo con centro más cercano. Sin embargo, difieren en los movimientos posteriores. Mientras el primero sugiere que se recalculen los centros al completar una iteración y se repita el procedimiento, el segundo establece que este re-cálculo se realice en cada movimiento de observaciones de un grupo a otro. Por su parte, Hartigan y Wong (1979) proponen elegir aleatoriamente los grupos iniciales y mover un individuo a otro grupo siempre que el movimiento minimice la suma de cuadrados dentro de los grupos, aún cuando esté en el grupo con centro más cercano. Para validar los agrupamientos puede usarse el estadístico *F*, resultante del cociente entre la sumas de cuadrados entre y dentro de los grupos conformados. Alternativamente, se pueden calcular ciertos índices de validación como el de Dunn (*di*), definido como la razón entre la mínima distancia entre grupos y la máxima distancia dentro de los mismos. Valores altos de este índice identifican grupos internamente densos y bien separados (DUNN, 1973).

La función **kmeans** del paquete **stats** permite obtener los agrupamientos resultantes de la aplicación del algoritmo *k* - medias. Esta función requiere al menos dos parámetros: la matriz de datos y la cantidad de centros iniciales a considerar y, permite obtener, entre otros resultados, las sumas de cuadrados entre y dentro de los grupos conformados. La función **cluster.stats**, del paquete **cluster**, brinda el índice de Dunn (*di)*.

Sucesivas experimentaciones, sobre el conjunto de datos “Iris de Fisher”, con diferentes métodos iniciales para *k*-medias proporcionan particiones semejantes entre sí. La estructura resultante de la aplicación del método propuesto por MacQueen (1967) es mostrado en el *clustplot* de la Figura 6 (F = 7.641194; di = 0.09880739), utilizando la misma notación que en la Figura 5. Los agrupamientos obtenidos son similares a los generados por *Complete Linkage*, *i.e*. los ejemplares setosa se agrupan de manera perfecta, evidenciándose confusión entre las especies virginica y versicolor.

Figura 6 – *Clustplot* de los 150 ejemplares de iris utilizando el criterio de *MacQueen* para k – medias (F = 7.641194; di = 0.09880739).



Fuente: Elaboración propia mediante R.

En la Figura 7 se incluyen los comandos implementados en **R** que permiten aplicar el método jerárquico aglomerativo *Average Linkage* y el método *k* - medias al conjunto de datos **iris**. El lector puede ejecutar el *script* para otros métodos aglomerativos modificando el segundo parámetro de la función **hclust**.

Figura 7 - Script de R para aplicar el método jerárquico *Average Linkage* y el método *k*-medias del conjunto “Iris de Fisher”

```
library(cluster)
d=daisy(iris[,1:4],metric="euclidean")
hc=hclust(d,method="average")
plot(hc,labels=iris[,5],hang=-1,cex=.7,xlab="Especímenes de iris")
dhc=cophenetic(hc)
cor(d,dhc)
k=3
grupos=cutree(hc,k)
rect.hclust(hc,k,border="red")
clusplot(iris[,1:4],grupos,color=TRUE,shade=FALSE,labels=4,lines=0,
          col.p=as.integer(iris[,5])+1,col.clus=c("blue","green","red"
    ))
table(iris[,5],grupos)
kmedias=kmeans(iris[,1:4],centers=k,algorithm="MacQueen")
kmedias$betweenss/kmedias$tot.withinss
kmedias$centers
library(fpc)
cluster.stats(d,clustering=kmedias$cluster)$dunn
clusplot(iris[,1:4],kmedias$cluster,color=T,shade=F,labels=4,lines=0,c
    ol.p=as.integer(iris[,5])+1,col.clus=c("blue","green","red"))
table(iris[,5],kmedias$cluster)
```

Fuente: Código propio.

Resultados del Análisis Lineal Discriminante

Para el ALD, una vez cargados los datos **Iris**, es necesaria la carga del paquete **MASS** y, entonces, se procede a utilizar la función **lda** para obtener la función discriminante lineal, la clasificación de los datos y la tasa de error para cada población. La función **lda** requiere como argumentos un factor (que en este trabajo identifica las especies) y las variables a considerar en la discriminación. Mediante el parámetro **CV** seteado a **TRUE**, la función realiza una validación cruzada que arroja la clase a la cual pertenecería cada elemento en la muestra y la probabilidad de pertenencia a cada clase o población. Si **CV** fuera seteado como **FALSE** (en este trabajo no fue incluida esta opción) devuelve las probabilidades *a priori* usadas; las medias de los grupos; y los coeficientes de la función lineal. El paquete **klaR** y la función **partimat** permiten realizar un gráfico, a partir de la elección de dos de las variables en el estudio, e identificar si las unidades de observación han sido clasificadas correctamente en la población de las que fueron observadas.

Así, por ejemplo, utilizando las cuatro variables Sepal.Length, Sepal.Width, Petal.Length y Petal.Width, la clasificación de las unidades elementales se muestra en la tabla de clasificación cruzada (Tabla 2). Se observa que los 50 datos de la población setosa están bien clasificados, mientras que, para la especie versicolor resultan dos datos mal clasificados, y uno se clasifica mal para la población virginica. Así, la tasa de error es: 0% para setosa; 4% para versicolor y 2% para virginica; lo cual muestra que la discriminación lineal es muy buena, pese a que no se verifican los supuestos de normalidad y homocedasticidad de los datos.

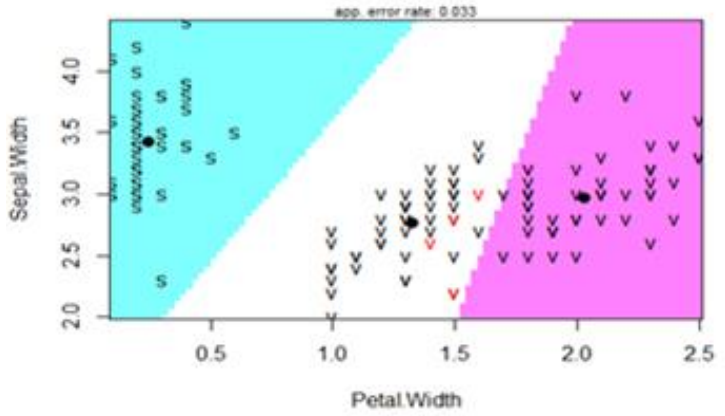
Tabla 2: Clasificación cruzadas para las 4 – variables y para cada especie de iris em estudio.

Clase	**Setosa**	**versicolor**	**Virginica**
Setosa	50	0	0
Versicolor	0	48	2
Virginica	0	1	49

Fuente: Resultados obtenidos mediante R.

Escogiendo las variables Sepal.Width y Petal.Width, en la Figura 8 se muestra la partición resultante y puede observarse que cuatro son los datos mal clasificados, o sea, cuatro datos de la especie virginica se clasificaron en la población versicolor. Representaciones equivalentes pueden obtenerse al seleccionar cualquiera de las seis posibles combinaciones de pares de variables.

Figura 8 – Diagrama de partición de los 150 ejemplares de iris escogiendo las variables ancho del sépalo (Sepal.Width) y del pétalo (Petal.Width).



Fuente: Elaboración propia mediante R.

Las sentencias incluidas en la Figura 9, ejecutadas en **R**, permiten realizar el Análisis Lineal Discriminante y obtener el diagrama de partición. Los comandos corresponden al ALD de las 4-variables y, las últimas dos líneas, al gráfico de partición de las variables Sepal.Length y Sepal.Width y al diagrama de dispersión de las 4-variables. El lector puede ejecutar los comandos para otras combinaciones de variables.

Figura 9 - Script de R para realizar el Análisis Lineal Discriminante y obtener el diagrama de partición del conjunto "Iris de Fisher"

```
f=5
if(is.factor(iris[,f])){pob=levels(iris[,f])
    npob=length(pob)}else{iris[,f]=as.factor(iris[,f])
        pob=levels(iris[,f])
    npob=length(pob)}
subdatos=1:4
library(MASS)
VC=TRUE
LinDis=lda(iris[,f]~iris[,1]+iris[,2]+iris[,3]+iris[,4],na.action="na.
     omit",CV=VC)
if(VC==TRUE){clase=LinDis$class}else{clase=predict(LinDis,iris[,subdat
     os])$class}
matriz=table(iris[,f], clase)
1-diag(prop.table(matriz,1))
1-sum(diag(prop.table(matriz)))
library(klaR)
partimat(x=iris[,1:2],grouping=iris[,f],method="lda",main="Gráfico   de
     partición")
plot(iris[,1:4],col=as.integer(iris[,f])+1,pch=19)
```

Fuente: Código propio.

Reflexiones y consideraciones finales

La mayoría de las técnicas del análisis exploratorio de datos multivariados requiere de dos supuestos básicos y esenciales para la aplicación de la teoría, que son la normalidad y homocedasticidad multivariada de los datos. El conjunto de datos "Iris de Fisher"

constituye uno de los ejemplos con datos reales más utilizados en la literatura del ADM, sin embargo, no satisface ninguno de los supuestos mencionados.

En este trabajo, se aplican dos técnicas de clasificación y discriminación multivariadas al conjunto "Iris de Fisher", mostrando resultados contradictorios con la teoría:

- el ALD, cuyo desarrollo teórico requiere de ambos supuestos, separa con probabilidades muy altas los elementos iris en las poblaciones y los clasifica de igual manera, en tanto,
- el AClu, cuyo desarrollo teórico no requiere de supuestos, no funciona bien con estos datos (los métodos sugieren, en general el uso de 2 grupos, pero cuando se usan 3, dado que se conoce que los individuos provienen de 3 poblaciones, los resultados no son los esperados).

Estas particularidades del conjunto "Iris de Fisher" lo constituyen en datos especialmente útiles para diseñar experiencias de aprendizaje que favorezcan la construcción de significados. La enseñanza basada en la resolución de problemas auténticos favorece un diseño instruccional apropiado para generar un aprendizaje significativo de las técnicas estadísticas aquí presentadas.

Las actividades descriptas en este trabajo para el abordaje de los procedimientos de discriminación y clasificación multivariadas, conciben el aprendizaje en el marco de la teoría del aprendizaje significativo, en tanto persiguen el objetivo de construir conocimientos que sean potencialmente transferibles a nuevas situaciones. El aprendizaje significativo, y constructivo en general, permite adquirir conocimientos estratégicos que se ponen en juego en la resolución de problemas auténticos, razón por la que resulta apropiado elegir esta estrategia de enseñanza. En este caso, se trata de obtener modelos adecuados para clasificar y discriminar especies de iris, siendo de suma importancia verificar ciertos supuestos requeridos por las técnicas aplicadas. La verificación de estos supuestos requiere que el estudiante recupere conocimientos previos no sólo para aplicarlos en un nuevo contexto, sino además para interpretar los nuevos conocimientos, asimilándolos e integrándolos a los anteriores. En ocasiones no es posible comprender los nuevos conocimientos pues las ideas previas son contrarias a ellos, tal como ocurre en este caso. Resulta necesario, entonces, generar un verdadero cambio conceptual, reorganizando toda la estructura de conocimientos. La resolución de problemas representa una estrategia potencialmente útil para promover este cambio, en tanto favorece prácticas reflexivas que apoyan la reestructuración conceptual.

Paralelamente, la obtención de un modelo adecuado, requiere de la utilización estratégica de ciertos conocimientos técnicos que hayan sido previamente ejercitados. Por esta razón, la propuesta didáctica incluye la resolución de una serie de ejercicios, más o menos repetitivos, que permiten adquirir tales destrezas. En síntesis, la propuesta didáctica presentada en este trabajo, está atravesada por el aprendizaje basado en problemas auténticos a fin de promover un aprendizaje significativo, en la que los ejercicios propuestos son necesarios para la adquisición de ciertas destrezas y conocimientos técnicos que sirvan como recursos para la resolución de los primeros.

Los procedimientos presentados en este trabajo pueden ser considerados como parte de una propuesta de aprendizaje para introducir los procedimientos del ALD y el AClu utilizando estos datos reales clásicos. Entendemos que, experiencias como ésta, permiten construir el conocimiento bajo el precepto de que la estadística es inseparable de sus aplicaciones, y su justificación final es su utilidad en la resolución de problemas externos a la propia estadística.

Referencias

ANDERSON, E. The irises of the Gaspé Peninsula. In: *Bulletin of the American Iris Society*, v. 59, p. 2–5, 1935.

AUSUBEL, D. *Psicología educativa*. México: Trillas, 1976.

BATANERO, C.; DÍAZ, C.; CONTRERAS, J.M.; ARTEAGA, P. Enseñanza de la estadística a través de proyectos. In: *BATANERO, C.; DÍAZ, C. (Eds.): Estadística con proyectos.* Granada: Departamento de Didáctica de la Matemática, p. 9–46, 2011.

BEZDEK, J. C.; KELLER, J.M.; KRISHNAPURAM, R.; KUNCHEVA, L.I.;PAL, N.R. Will the Real Iris Data Please Stand Up? In: *IEEE Transactions on Fuzzy Systems*, v. 7, n. 3, p. 368–369, 1999.

DE OLIVEIRA BUSSAB, W.; MIAZAKI, E. S.; de ANDRADE, D. *Introdução à Análise de Agrupamentos.* San Pablo: Asociación Brasilera de Estadística, 1996.

DÍAZ BARRIGA, F. Cognición situada y estrategias para el aprendizaje significativo. In: *Revista Electrónica de Investigación Educativa*, v. 5, n. 2, 2003.

DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. In: *Journal of Cybernetics,* v. 3, n. 3, p. 32–57, 1973.

FISHER, R. A. The Use of Multiple Measurements in Taxonomic Problems. In: *Annals of Eugenics,* v. 7, n. 2, p. 179–188, 1936.

FORGY, E. W. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. In: *Biometrics*, v. 21, p. 768–769, 1965.

FRANKLIN, C.; KADER, G.; MEWBORN, D.; MORENO, J.; PECK, R.; PERRY, M.; SCHEAFFER, R. *Guidelines for assessment and instruction in statistics education (GAISE) report: A Pre-K-12 curriculum framework.* Alexandria, VA: American Statistical Association, 2005.

GAL, I. Adult's statistical literacy. Meanings, components, responsibilities. In: *International Statistical Review*, v. 70, n. 1, p. 1–25, 2002.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. In: *Applied Statistics*, v. 28, p. 100–108, 1979.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 6th ed. New Jersey: Pearson Prentice Hall, 2007.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley, 1990.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: L. M. Le Cam y J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, p. 281–297, 1967.

PISON, G.; STRUYF, A.; ROUSSEEUW, P. J. Displaying a clustering with CLUSPLOT. In: *Computational Statistics & Data Analysis***,** v. 30, p. 381–392, 1999.

POZO MUNICIO, J. I.; PÉREZ ECHEVERRÍA, M. del P. Aprender para comprender y resolver problemas. In: *(Eds.), Psicología del aprendizaje universitario*. Madrid: Ediciones Morata, p. 31–53, 2009.

SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. In: **Taxon**, v. 11, p. 33–40, 1962.

WELCH, B. L. Note on Discriminant Functions. In: *Biometrika*, v. 31, p. 218–220, 1939.

WILD, C.; PFANNKUCH, M. Statistical thinking in empirical enquiry. In: *International Statistical Review*, v. 67, n. 3, p. 223–265, 1999.

Texto recebido: 20/04/2018
Texto aprovado: 10/04/2019