Valores atípicos en los datos, ¿cómo identificarlos y manejarlos?

Outliers in data sets, how identify and handling them?

Leneidy Pérez Pelea[1]

RESUMEN
En el análisis de datos experimentales, es frecuente encontrar variables biológicas con distribución no normal, en las cuales no se cumplen también, otras de las premisas planteadas en los métodos estadísticos tradicionales. En ocasiones, la falta de normalidad puede atribuirse a la presencia de uno o más valores atípicos (*outliers)* en los datos, los cuales se desvían mucho del resto de los valores y caen fuera del patrón general de distribución de la variable. Varios autores han propuesto diferentes definiciones para estos valores y han desarrollado métodos muy variados para identificarlos y manejarlos. Los métodos más empleados están basados en análisis de distancia, agrupamientos, varianza, ángulos entre vectores y densidad en la vecindad de las observaciones. También varían en depencia de si los valores atípicos están presentes en experimentos que analizan una o múltiples variables. Existe una gran controversia en la literatura en relación con la eliminación de los valores atípicos. Se ha planteado que se debe conocer su causa y la influencia que pueden tener en los resultados de los experimentos, antes de tomar la decisión de eliminarlos o incluirlos en el análisis, porque cambian las inferencias que se obtienen y, en ocasiones, su eliminación puede conducir a la pérdida de una información importante. En el presente artículo se hace una revisión de las principales causas que pueden provocar la aparición de estos valores atípicos, y algunos de los métodos que se han propuesto para identificarlos y manejarlos.

Palabras clave: valores extremos, pruebas de detección de anomalías, pruebas de discordancia

ABSTRACT
In experimental data analysis it is frequently found no normal biological variables, in which other assumptions of traditional statistics methods are violated. Sometimes, no normally is due to the presence of one or more outlier values, which are far away the other values and fall out the general patron of the variable distribution. Different definitions for this values were proposed by various authors, who also have developed a variety of methods to identify and handle outliers. The most employ methods are based on distance, clustering, variance, angle between vectors and density in the neighborhood of the observations. These methods are also different when there are one or more variables in the experiment. There are a great controversy on the literature related with the elimination of outliers. The cause of outlier and its influence on the results of experiments should be known before taking the decision of its elimination, because the outliers change the inferences of the experiment, and sometimes, its elimination can lead up to the loss of important information. In the present paper, it was made a revision about the main causes of outliers and some proposed methods to identify and handle them.

Keywords: outliers, anomalies detections tests, discordant tests

Recibido: mayo 2019 **Aceptado**: noviembre 2019
Publicado online 31 de diciembre de 2019. ISSN 2410-5546 RNPS 2372 (DIGITAL) - ISSN 0253-5696 RNPS 0060 (IMPRESA)

INTRODUCCIÓN

En las investigaciones que se realizan en el campo de las ciencias biológicas y agrícolas, los investigadores utilizan con frecuencia métodos de análisis estadístico, que requieren del cumplimiento de premisas como el ajuste de los datos a la normalidad, la homogeneidad de varianzas, la independencia de errores experimentales, entre otras, las cuales fueron especificadas por el autor de la prueba en cuestión. Estos métodos paramétricos son de preferencia por parte de los investigadores, sobre los métodos no paramétricos porque tienen una potencia superior, siempre que el tamaño de la muestra sea suficientemente grande y se verifique el cumplimiento de las premisas establecidas (Pérez 2018). El empleo de métodos estadísticos paramétricos cuando no se cumple alguna de estas premisas, conlleva a conclusiones no válidas y a errorres en el proceso de toma de decisiones (Herrera & *al.* 2012).

La detección de una distribución no normal o de varianzas heterogéneas en los datos, es producto en algunos casos, de la presencia de uno o más valores que se encuentran muy alejados del resto de los valores de la muestra, los cuales se denominan observaciones atípicas, aberrantes, inconsistentes o muy extremas (*outliers* en inglés) (Whitlock & Schluter 2009; Zar 2010). Estas observaciones, aunque tienen características diferentes al resto, no pueden ser categorizadas como beneficiosas o problemáticas, sino que deben ser contempladas en el contexto del análisis y evaluarse el tipo de información que pueden proporcionar. Su principal problema radica en que pueden ser elementos no representativos de la población, por lo que su presencia en la muestra sesgará los resultados de las pruebas estadísticas tradicionales (Garson 2012; Kwak & Kim 2017). Estos valores pueden ser inevitables, especialmente en grupos grandes de datos (García 2017).

[1]Departamento de Biología Vegetal, Facultad de Biología, Universidad de la Habana. Calle 25 # 455, e/ I y J Vedado, La Habana, Cuba. C.P. 10400. e-mail: lene@fbio.uh.cu
Editor encargado: José Angel García-Beltrán.

Existe un gran debate en la literatura con relación a si los *outliers* deben ser eliminados o no, por la influencia que tienen en los resultados de los análisis. Algunos autores están a favor de su eliminación porque consideran que es honesta, deseable e importante (Osborne 2002; Judd & *al*. 2009). Un *outlier* puede ser resultado de una unidad experimental que no pertenece a la población en estudio, es un individuo de otra especie, tiene una edad muy diferente al resto de las unidades experimentales de la muestra, o puede ser producto de errores que se cometieron durante la conducción del experimento. Zar (2010) planteó que cuando el investigador está seguro que el *outlier* es producto de alguna de estas causas, lo puede eliminar. Sin embargo, no siempre resulta correcta la eliminación de estas observaciones. Algunos *outliers* pueden ser fenómenos interesantes que conllevan al descubrimiento de un conocimiento inesperado u observaciones correctas obtenidas por azar en la población, por lo que su eliminación puede conducir a la pérdida de información útil y a la obtención de resultados inexactos o incorrectos en los análisis estadísticos.

El término *outlier* a veces es empleado solo para hacer referencia a los valores externos, no extremos. Un valor externo no forma parte de la población de interés por lo que sin lugar a dudas debe ser eliminado, mientras que un valor extremo si representa una fracción de esta, y aún cuando puede perjudicar el empleo de uno u otro método, debe incluirse en el análisis.

DEFINICIÓN Y CLASIFICACIÓN DE LOS *OUTLIERS*

Los *outliers* han sido considerados un serio problema para la aplicación de muchos procedimientos estadísticos, especialmente aquellos que requieren el cumplimiento de la premisa de normalidad. Varios autores han propuesto definiciones del término y han desarrollado procedimientos para su identificación y manejo. En una amplia revisión de literatura científica realizada por Aguinis & *al.* (2013), estos autores encontraron 14 definiciones de *outliers*, 39 técnicas de identificación y 20 técnicas de manejo, las cuales resumen en tablas que deben ser consultadas por los investigadores. Dado el gran número de definiciones y técnicas descritas, no resulta sorprendente que en algunos artículos y textos, se brinden recomendaciones inconsistentes sobre cómo definir, identificar y manejar los *outliers*. Además, algunos investigadores como Kulich & *al.* (2011) han utilizado un método de identificación de *outliers* que es inconsistente con la forma en que los definen.

Hawkins (1980) definió un *outlier* como una observación que por ser muy diferente a las otras observaciones de un mismo conjunto de datos, puede considerarse que fue creada por un mecanismo diferente. Barnett & Lewis (1994) lo definieron como una o varias observaciones del grupo de datos que se desvían marcadamente del resto de los miembros de la muestra. De manera similar, Johnson & Wichern (1992) plantearon que un *outlier* es una observación del grupo de datos que parece ser inconsistente con el resto de las observaciones y Kwak & Kim (2017) definen los *outliers* como valores extremos que caen anormalmente por fuera del patrón general de distribución de la variable.

Los *outliers* pueden ser simples o multivariados. Cuando los valores son atípicos con respecto a una sola variable se denominan *outliers* simples. Los *outliers* multivariados son valores extremos con respecto a múltiples variables (Garson 2012). Si los datos contienen un número significativo de *outliers*, se deben utilizar técnicas estadísticas robustas para su análisis, que no son sensibles a la presencia de estos valores.

Zhang (2013) propuso que los *outliers* se podían clasificar en puntuales (*point outliers*) y colectivos (*collective outliers*), basado en el número de datos involucrados en el concepto. Los puntuales son valores individuales que se encuentran muy alejados del resto de las observaciones y han sido el interés de la mayoría de las pruebas de detección, pues son el tipo más simple. Los colectivos representan una colección de valores que se encuentran muy alejados del resto de las observaciones. Los valores individuales que forman un *outlier* colectivo no se consideran *outliers* por sí mismos, sino que es el conjunto de valores el que se considera anómalo o atípico. Sin embargo, aunque son muchas las definiciones que se han dado sobre este concepto, lo que caracteriza una observación como atípica es el impacto que produce en los resultados que se obtienen cuando se analizan los datos.

PRINCIPALES CAUSAS Y CONSECUENCIAS DE LOS VALORES ATÍPICOS

Los *outliers* se pueden organizar en dos categorías: los que son producto de errores en los datos y los que se deben a la variabilidad inherente en estos. Dan & Ijeoma (2013) resumieron varias de las posibles causas de la presencia de *outliers* en una muestra y cómo se pueden manejar. Según estos autores, los *outliers* pueden ser resultado de:

(1) Errores humanos en la colecta, registro o entrada de los datos, los cuales pueden corregirse revisando los documentos o los sujetos experimentales si es posible. También es importante revisar cuidadosamente la base de datos antes de correr los análisis en un programa estadístico, pues se pueden detectar errores tipográficos u omisión de valores, que alteren los resultados.

(2) Información brindada incorrectamente a los investigadores o encuestadores de manera intencionada, cuando las preguntas o temas que se tratan pueden

resultar sensibles a los sujetos experimentales y estos no dan la información correcta.

(3) Errores cometidos en el muestreo, como falta de aleatorización o la selección de individuos que pertenecen a otra población, cuyos datos no van a reflejar el comportamiento de la población objeto de estudio; los cuales se pueden eliminar.

(4) Fallos causados por la metodología de la investigación, el comportamiento anómalo de algún sujeto experimental, fenómenos no usuales que pueden ocurrir dentro o fuera de los laboratorios, fallo o mala calibración de los equipos; los cuales pueden eliminarse si el investigador no está interesado en estudiar el fenómeno particular en cuestión.

(5) Premisas incorrectas acerca de la distribución de los datos debido a que en algunas variables se pueden observar distribuciones asimétricas, bimodales, asintóticas o aplastadas. Algunos datos tienen una estructura diferente a la que originalmente asumió el investigador. En dependencia de los objetivos de la investigación, los *outliers* pueden o no representar la variabilidad inherente en los datos y se deben mantener en los análisis.

(6) Unidades experimentales muestreadas en la población correcta y que se obtienen por azar. En estos casos, el tamaño de la muestra juega un rol importante en la probabilidad de obtener *outliers*. Cuando se trabaja con muestras grandes se requieren métodos más sofisticados para la detección de los *outliers*, pues su número se incrementa con el aumento del tamaño de la muestra (Cleophas & Zwinderman 2019).

Evans (1999) planteó que en una población normalmente distribuida, es más probable que un valor dado provenga del área más concentrada de la distribución, que de las colas o extremos. Existe alrededor de un 1% de posibilidad de obtener un *outlier* en una población normalmente distribuida, o sea, en promedio alrededor del 1% de los datos estará a tres unidades de desviación estándar de la media. Cuando los *outliers* son producto a la variabilidad inherente en los datos, las opiniones sobre cómo proceder difieren ampliamente. Debido a los efectos que tienen los *outliers* sobre la potencia, la exactitud y las tasas de error experimental, es necesario usar alguna estrategia de transformación de escala para mantener ese valor en el grupo de datos y minimizar los daños en la inferencia estadística (Osborne 2002).

(7) Focos potenciales de investigación: los *outliers* pueden representar errores molestos, datos legítimos o pueden inspirar la investigación. Por ejemplo: en África se observó que algunas mujeres llevaban años viviendo con VIH sin tratamiento, las cuales constituyen casos atípicos cuando se comparan con la mayoría de las mujeres no tratadas que mueren rápidamente. Esta información pudiera ser descartada como ruido o error, pero puede servir como fuente de inspiración para investigar esta situación. Antes de descartar los *outliers*, los investigadores necesitan considerar si estos datos contienen información valiosa, que no está necesariamente relacionada con el estudio, pero que tiene importancia en un sentido más global.

Los *outliers* se pueden observar también en estudios de recalibración de los laboratorios. Estos valores se pueden obtener a través de procesos no relacionados con los procedimientos de mediciones en los laboratorios, como mezclas inadecuadas, evaporación, degradación, marcajes incorrectos o errores en la entrada de los datos, los cuales no son informativos acerca de la recalibración de la mayoría de las muestras, donde no ocurren estos errores. La presencia de estos valores disminuye la precisión y aumenta los errores en la recalibración, por lo que se hace importante hacer una distinción entre los errores relacionados y no relacionados con el proceso de recalibración. Se han descrito varios enfoques para identificar y minimizar el efecto de los *outliers*, cuando se comparan mediciones en estudios de recalibración en los laboratorios. Parrinello & *al.* (2016) presentaron un método iterativo de identificación y eliminación de *outliers*, en estudios de recalibración en laboratorios.

¿Qué debe hacerse si en la investigación no se encuentra una causa probable de la presencia del *outlier*? Murphy y Lau (2008) propusieron realizar un análisis de los datos con el *outlier* y sin él; si las conclusiones son diferentes, entonces se considera que el valor tiene influencia y esto debería indicarse en los resultados. Otra opción es utilizar estimadores robustos para caracterizar los grupos de datos, tal como la mediana de la muestra en lugar de la media, que no es sensible a los *outliers*. El método de detección de *outliers* basado en la mediana se considera robusto y es menos afectado por el tamaño de la muestra que otros métodos tradicionales de detección (Sandbhor & Chaphalkar 2019).

La presencia de *outliers* causa efectos negativos en el análisis de los datos. Osborne & Overbay (2004) plantearon que los *outliers* generalmente incrementan la varianza del error, reducen la potencia de la prueba estadística, sesgan la estimación de parámetros, afectan el ajuste a la distribución normal, influyen en el no cumplimiento de las premisas de esfericidad y normalidad multivariada en análisis multivariados y alteran la relación entre los errores tipo I y II. Además, incrementan las tasas de error experimental y distorsionan sustancialmente los estimados de los parámetros y los estadísticos, cuando se utilizan pruebas paramétricas y no paramétricas (Dan & Ijeoma 2013).

Rara vez en los artículos científicos, los autores plantean que hicieron una búsqueda de posibles *outliers*. Osborne & *al.* (2001) encontraron que solo en el 8% de los artículos revisados, los autores indicaron que verificaron el cumplimiento de las premisas del procedimiento estadístico empleado, incluida la búsqueda de *outliers*. Este resultado es alarmante si se tiene en consideración la influencia que tienen estos valores en las inferencias que se realizan.

La presencia de estas observaciones en un conjunto de datos, conduce a cambios importantes en la estimación de parámetros cuando se utilizan métodos estadísticos que emplean estimadores de máxima verosimilitud (Aguinis & *al.* 2013). Para poder solucionar este problema, se deben encontrar métodos que permitan la identificación y el manejo de dichas observaciones.

MÉTODOS DESARROLLADOS PARA LA DETECCIÓN Y EL MANEJO DE LOS *OUTLIERS*

Entre los métodos más empleados para la detección de valores atípicos en datos con baja dimensionalidad, se encuentran: métodos estadísticos, métodos basados en distancia, en densidad y en agrupamientos (Zhang 2013). Otros autores como Aggarwal (2013) y De Armas (2015), propusieron además, métodos basados en profundidad, en varianzas y en ángulos entre los vectores. También se han utilizado las redes neuronales para identificar *outliers* en tareas de clasificación o regresión (Sykacek 1997). Algunas técnicas de la minería de datos se han enfocado en la detección de valores atípicos en grandes bases de datos, a través de diferentes algoritmos que han evolucionado en términos de efectividad y eficiencia (De Armas 2015).

La detección y el manejo de los *outliers,* en datos numéricos muestrales, es una parte importante de muchos de los procesos de estimación (Van der Loo 2010). El primer paso para manejar los *outliers*, es encontrarlos, pues deben ser identificados antes de decidir si se eliminan o no. Generalmente, los usuarios modelan los grupos de datos con el empleo de una distribución estadística, y los puntos se determinan como *outliers* en dependencia de como aparecen en relación con el modelo postulado. El principal problema con esta técnica, es que en ocasiones, el usuario puede no tener suficiente conocimiento sobre la distribución muestral de los datos (Pamula & *al*. 2011).

Las técnicas propuestas para la detección de *outliers* usan algoritmos y mecanismos diferentes y dependen de las características de los datos (Zhang 2013). Los algoritmos que ayudan a detectar *outliers*, identifican automáticamente si la observación es valiosa o no, en grandes colecciones de datos. Cada algoritmo está basado en un modelo que confía en ciertas premisas de cuál valor califica como *outlier*, y la aplicabilidad de cada modelo depende de la naturaleza de los datos. Aggarwal (2013) planteó que virtualmente todos los algoritmos de detección de *outliers* crean un modelo del patrón normal de los datos, para calcular una puntuación como *outlier* en base a la desviación de ese patrón, o sea, la puntuación del dato dado se calcula al evaluar la calidad de ajuste del punto al modelo.

Hasta el presente han sido descritos un número considerable de métodos que permiten detectar la presencia de *outliers* en análisis univariados y multivariados. En un artículo publicado por Aguinis & *al.* (2013) se resumen en una tabla, 39 técnicas de identificación, que van desde métodos gráficos, como: gráficos de caja y bigote, de dispersión, de rama y hoja (*Stem and leaf plot*), de cuartiles (*Q-Q plot*) y probabilidades (*P-P plot*); a técnicas basadas en medidas de desviación estándar, residuos estandarizados o estudentizados, métodos estadísticos paramétricos y no paramétricos, basados en medidas de distancias, de agrupamiento, bondad de ajuste, entre otros. Algunos de estos métodos se describen a continuación.

Los métodos de detección estadísticos se pueden clasificar en paramétricos y no paramétricos. En los paramétricos se asume la distribución subyacente a los datos y en los no paramétricos no se asume algún conocimiento de las características de la distribución. Teniendo en cuenta la distribución que se asume que ajusta a los datos, se consideran *outliers* a aquellos puntos que no se ajustan con el modelo subyacente (Zhang 2013). Una de las ventajas principales de estos modelos es que pueden ser aplicados en cualquier tipo de datos, siempre que se pueda encontrar un modelo generativo apropiado para cada componente (De Armas 2015). Su desventaja es que intentan ajustarse a una distribución particular que no siempre puede ser apropiada para los datos, por lo que se debe tener un buen conocimiento del proceso generativo de los datos pues la elección del modelo influye sobre los resultados obtenidos (Aggarwal 2013). Además, no se recomienda su aplicación en datos multidimensionales porque resulta difícil conocer la distribución subyacente de los datos.

Los métodos paramétricos incluyen los análisis regresión que se emplean mayormente en datos de series de tiempo y los basados en la distribución Gaussiana que utilizan estimadores de máxima verosimilitud, entre los que se encuentra la prueba media-varianza (Método *Z-score*) y los gráficos de caja y bigote. Entre los métodos no paramétricos se encuentran el uso de histogramas y de métodos de función de Kernel para aproximar la verdadera función de densidad probabilística (Zhang 2013).

Los modelos de regresión pueden ser lineales o no lineales, según el ajuste de los datos. Un valor se considera atípico si hay una marcada desviación entre el valor y su valor esperado cuando se realiza el ajuste por regresión, o sea, cuando tienen errores o residuos muy grandes. El método del valor de Z (*Z-score*) utiliza los valores de la media y la desviación estándar para calcular el valor de Z (distribución normal estandarizada) en datos con distribución normal. El valor de Z indica a cuántas unidades de desviación estándar se encuentra un valor de la media, por lo que, si para una observación dada, el valor de Z calculado es modularmente mayor que tres, esta se considera un posible *outlier* (Aggarwal 2013). Este método tiene algunas limitantes porque la media y la desviación estándar son sensibles a la presencia de *outliers* en la muestra, lo que causa un problema de enmascaramiento, que impide la detección de los *outliers* menos extremos por causa de los más alejados (Seo 2006). Esta prueba se puede utilizar cuando los datos se ajustan a otras distribuciones como t de Student y Poisson (Zhang 2013).

Como la media y la desviación estándar son estadísticos sensibles a la presencia de *outliers*, la prueba no se considera apropiada. De manera alternativa, la mediana y el rango intercuartil, que son estadísticos descriptivos no sensibles a la presencia de valores atípicos, se utilizan para construir gráficos de caja y bigote. Cualquier valor que se encuentre por fuera de los límites de los bigotes del gráfico, se considera atípico y se representa con un punto o un asterisco (Kwak & Kim 2017).

Existe una modificación del método *Z*-score, que utiliza la mediana (Md) y la desviación absoluta de la mediana (*Median Absolute Deviation*, *MAD*) en lugar de la media y la desviación estándar, que es un método básico robusto no sensible a la presencia de *outliers* (Iglewicz & Hoaglin 1993). El valor de MAD es un estimador de la dispersión de los datos, similar a la desviación estándar. En datos con distribución normal, se calcula un valor de Z modificado (Mi) con el empleo de la siguiente fórmula: Mi = 0,6745 (xi – Md) / MAD. Iglewicz & Hoaglin (1993) sugirieron que las observaciones con valores modulares de Mi mayores que 3,5 se pueden identificar como *outliers.*

Los histogramas constituyen la mejor manera de comprobar la presencia de valores atípicos en datos univariados de variables continuas. El histograma divide un rango de valores en varios grupos y se muestra su frecuencia con una barra. Si se asume que estos grupos están arreglados en un orden ascendente, se pueden detectar valores muy pequeños o muy grandes en los extremos (García 2017). Otro método no paramétrico utiliza la función de Kernel para aproximar la función de densidad probabilística, se consideran atípicos aquellos valores que caen en el área de baja probabilidad de la función de densidad. Este método ha sido empleado en disímiles campos como: análisis de imágenes, detección de intrusión y de sensores en redes informáticas, y se puede aplicar en datos univariados y multivariados (Zhang 2013).

En los métodos basados en distancias, los *outliers* se detectan en base a una medida de distancia, que se calcula con todas las dimensiones posibles entre un punto y su vecindario dentro del conjunto de datos. Se utilizan diferentes medidas de distancia como la Euclideana, la de Manhattan y la de Mahalanobis. Knorr & Ng (1998) fueron los primeros en proponer un método de detección de *outliers* basado en distancias, el cual fue muy bien aceptado porque generalizaba varias pruebas estadísticas de detección. Otros investigadores como Ramaswamy & *al.* (2000) y Angiully & *al.* (2006) propusieron extensiones de este método. Un método para detectar *outliers* basado en distancia local fue presentado por Zhang & *al.* (2009), en el cual se determina el grado en el cual una observación se desvía de su vecindario, al calcular el factor de *outlier* basado en distancia local (*Local Distance-based Outlier Factor*, *LDOF*).

Pamula & *al.* (2011) propusieron un método en el cual se identifican los puntos que no son *outliers* con el empleo de las funciones de agrupamiento y distancia, y se separan del resto. Se usa un algoritmo de k-medias para agrupar los datos y se separan los puntos cuya distancia del centroide del grupo es menor que el radio del respectivo grupo, y para cada punto no separado en cada grupo, se calcula la medida *LDOF*. Se reportan como *outlier*s los puntos con los mayores valores de *LDOF*.

La principal ventaja de los métodos basados en distancia es que son no paramétricos, por lo que no asumen una distribución que ajuste los datos. Además, son técnicas fáciles de comprender e implementar. Su mayor desvantaja es que no son efectivos en datos con alta dimensionalidad, porque las definiciones de vecindario local o vecinos más cercanos, no tienen mucho sentido en un espacio de alta dimensionalidad (Zhang 2013).

Las técnicas basadas en densidad toman en consideración la densidad de los datos, cuando se calculan las distancias entre los puntos del conjunto de datos, para detectar como *outliers* locales a aquellos que se encuentran en zonas de baja densidad (Breunig & *al.* 2000). Estos autores definieron el concepto de Factor de *Outlier* Local (*Local Outlier Factor, LOF*), que fue el primero en cuantificar cuan alejado se encuentra el *outlier*. El factor es local porque solo tiene en cuenta un vecindario restringido por los vecinos más cercanos, para cada observación. El valor *LOF* de un objeto se obtiene al comparar su densidad con la de su vecindario, por lo

que su capacidad de modelar es más fuerte que cuando se considera la densidad del objeto solo. Con relación a los métodos basados en distancia, las técnicas basadas en densidad son más efectivas, pero son más complejas y costosas computacionalmente (Zhang 2013).

Los métodos basados en agrupamientos definen los *outliers* como puntos que se localizan fuera de cualquiera de los grupos (*clusters*) generados por algoritmos de agrupamiento. Existen muchos estudios relacionados con las técnicas de agrupamiento y algunas de ellas están equipadas con mecanismos para reducir los efectos adversos de los *outliers* (Zhang 2013).

En los métodos basados en profundidad se utiliza el polígono mínimo convexo, que es la región convexa más pequeña que contiene a todos los puntos, para encontrar los *outliers*. Las esquinas del polígono convexo contienen los límites exteriores de los datos. En este método, los datos son organizados en capas en el espacio, con la expectativa de que las capas más superficiales contendrán los datos *outliers* con mayor probabilidad que las capas más profundas. No se tienen que ajustar los datos a una distribución específica y se pueden procesar datos multidimensionales. El método tiene un algoritmo iterativo, con una complejidad computacional que se incrementa exponencialmente con la dimensión de los datos, lo que lo hace impráctico e ineficaz en dimensionalidades muy grandes (Aggarwal 2013).

Los modelos basados en varianzas miden el impacto de los *outliers* en la varianza de los datos. La premisa de base es que el *outlier* se ubica en los límites de los datos, por lo que al eliminarlo se reducirá la varianza de manera significativa. Los *outliers* son definidos como un conjunto de puntos cuya eliminación causa la máxima reducción en la varianza de la muestra. Este enfoque es independiente de la distribución y puede ser aplicado a cualquier conjunto de datos (Aggarwal 2013).

Existen otras pruebas de detección de *outliers*, frecuentemente llamadas pruebas de discordancia, muchas de las cuales está basadas en pruebas estadísticas que tienen en cuenta la distancia de una observación a la localización del parámetro y su dispersión en la muestra, como la Prueba de Dixon y la Prueba de Grubbs (Van der Loo 2010). Estas técnicas están diseñadas para detectar un único *outlier* en un grupo de datos, y por lo tanto no son adecuadas para la detección de múltiples *outliers*. Cuando hay múltiples *outliers* en un grupo de datos, la investigación resulta más complicada, porque pueden presentarse los fenómenos de enmascaramiento (*masking*) y empantanamiento (*swamping*). Un *outlier* enmascara a otro, cuando este último es considerado atípico solo por sí mismo, pero no en presencia del primero. Por el contrario, un *outlier* empantana a otro, si este último valor solo se considera atípico bajo la presencia del primero (Muñoz & Amón 2013). Resulta importante destacar que lo primero es considerar los datos gráficamente para identificar la posible existencia de más de un *outlier*, ya sea en la misma dirección o en la dirección opuesta, antes de utilizar alguna de estas técnicas y verificar que los datos se ajustan a la distribución normal (Murphy & Lau 2008). A diferencia de las pruebas anteriores, la Prueba de Rosner permite detectar múltiples *outliers* en un grupo de datos que contenga al menos 25 observaciones normalmente distribuidas.

Van der Loo (2010) propuso un método en el cual los *outliers* son definidos como observaciones que no son probablemente generadas por la distribución de la mayoría de los datos. Después de obtener un estimado robusto para la distribución de la mayoría, se conciben dos pruebas estadísticas. La primera es el valor de los datos sin transformar y la segunda es la regresión residual usada en el procedimiento de estimación robusto. La principal ventaja de este método es que brinda resultados muy robustos una vez que se encuentre la distribución correcta. El método ha sido desarrollado para las distribuciones Weibull, Pareto, exponencial, lognormal y normal con el empleo de la plataforma de análisis de datos R.

Las pruebas estadísticas que consideran las relaciones entre variables diferentes son esenciales para detectar *outliers* multivariados (Kwak & Kim 2017). El análisis de *outliers* también se puede aplicar a datos multivariados en diversas formas, algunas de las cuales tratan de modelar la distribución subyacente de forma explícita, mientras que otras se basan en un análisis estadístico más general, que no asume ninguna distribución particular. Los *outliers* tienen un impacto dramático sobre los resultados de los análisis estadísticos multivariados, por ejemplo: pueden distorsionar los coeficientes de correlación al obtenerse estimados sesgados e influyen en la linealidad entre las variables; pueden conducir a problemas de colinealidad entre las variables independientes de un análisis de regresión múltiple; influyen en la formación de los grupos en el análisis de conglomerados, porque pueden constituir el centro de un grupo o el punto de partida (Finch 2012).

Los *outliers* multivariados no pueden ser detectados cuando cada variable se considera de forma independiente, sino que deben observarse las interacciones entre las diferentes variables (Muñoz & Amón 2013). El enfoque más comúnmente recomendado para detectar *outliers* multivariados está basado en una distancia multivariada, la distancia de Mahalanobis, la cual se ha utilizado en múltiples campos. Su utilidad radica en que es una forma de determinar la similitud entre dos

variables aleatorias multidimensionales. Se diferencia de la distancia Euclideana en que tiene en cuenta la correlación entre las variables aleatorias.

A pesar de las ventajas que tiene su empleo, este valor de distancia sensible a los *outliers* porque está basado en la matriz de covarianzas muestral, que es también sensible a los *outliers* (Wilcox 2005). Además, se calcula para variables continuas, por lo que resulta inapropiado cuando los datos son categóricos. Por otra parte, cuando se emplea la distancia de Mahalanobis como criterio para la detección de valores atípicos, pueden influir los efectos de enmascaramiento que disminuyen la distancia de un valor típico o de empantanamiento, que aumentan la distancia de observaiones que no son *outliers*. Estos problemas pueden resolverse con el uso de estimadores robustos que son menos afectados por los *outliers* (Muñoz & Amón 2013). Por estas razones, los investigadores han desarrollado alternativas para la detección de *outliers* multivariados, que son más robustas y flexibles.

Una alternativa utilizada para la detección de *outliers* multivariados es el método del Elipsoide de Mínimo Volumen *(Minimum Volume Ellipsoid, MVE*) (Rousseeuw & Leroy 2003). El objetivo de este método es identificar una submuestra de observaciones, que crea la elipsoide de menor volumen de los datos, basada en los valores de todas las variables. Por definición, esta elipsoide debe estar libre de *outliers* y a partir de este subgrupo de observaciones se deben obtener los estimados de tendencia central y dispersión (Finch 2012). Otra alternativa es el método del Mínimo Determinante de Covarianza (*Minimum Covariance Determinant, MCD*), el cual busca una porción de los datos que elimina la presencia y el impacto de los *outliers*, al minimizar el determinante de la matriz de covarianza, que es un estimado de la varianza generalizada en un grupo de datos multivariados (Rousseeuw & van Driessen 1999).

Los dos métodos descritos anteriormente (*MVE* y *MCD*) tienden a identificar un número grande de *outliers*, cuando las variables que se examinan no son independientes entre sí (Wilcox 2005). Para evitar este problema, se diseñó el método de Mínima Varianza Generalizada (*Minimum Genralized Variance, MGV*), en el cual se selecciona el grupo con la menor varianza. Este método también usa un algoritmo iterativo. Se consideran *outliers* aquellos puntos con varianza generalizada superior a un valor, determinado a partir de la mediana y los cuartiles inferior y superior de los valores de varianza generalizada, cuya fórmula de cálculo varía en función del número de variables analizadas (Ver Finch 2012).

En la actualidad existen herramientas informáticas que permiten realizar el análisis de grandes cantidades de información, lo que posibilita identificar de manera más sencilla, las correlaciones entre las variables y la presencia de *outliers* multivariados. La plataforma de análisis de datos R y el programa estadístico SPSS se encuentran entre los más utilizados para el análisis mutlivariado de los datos, y presentan comandos y métodos gráficos para la detección de *outliers* multivariados (Muñoz & Amón 2013).

También, Aguinis & *al.* (2013) resumieron en una tabla varias técnicas descritas en la literatura para manejar la presencia de *outliers*. Entre las técnicas descritas se observan métodos sencillos como: corregir o modificar los valores incorrectos; eliminar los valores incorrectos; estudiar el *outlier* como un fenómeno único de interés; aplicar determinadas funciones matemáticas para transformar todos los datos con vistas reducir la varianza del error y disminuir la asimetría de los datos. También se pueden transformar los valores extremos a un percentil específico de los datos (*Winsorization*), de modo que el percentil 90 *winsorizado* transforma todos los datos por debajo del percentil 5 a este y todos los datos por encima del percentil 95 a este. Existen otras técnicas como: el método de mínimos cuadrados recortados (*trimmed*) en el cual se ordenan ascendentemente los cuadrados de los residuos de cada caso y se recortan o remueven los mayores valores; y los métodos no paramétricos que trabajan con rangos y están menos influenciados por los *outliers* que los valores originales.

Estos autores también propusieron de la literatura consultada, otros métodos robustos más complejos que se pueden emplear para analizar los datos con *outliers,* entre los que se encuentran: la M-estimación que se utiliza en las series de tiempo, en la cual se reduce el efecto de la influencia de los *outliers* al reemplazar los residuos cuadráticos por otra función de los residuos. Se pueden emplear estadísticos Bayesianos que acercan los *outliers* al centro o al centroide de los datos. También se ha descrito el uso de la distancia de Mahalanobis para asignar ponderaciones a cada punto, de manera que los casos que son extremos en las variables predictoras son ponderados hacia abajo en un procedimiento en dos etapas, u otra alternativa en la cual la ponderación es asignada a través de un algoritmo iterativo de mínimos cuadrados reponderados. Los métodos de ecuaciones de estimación generalizadas se pueden utilizar para limitar la influencia de los *outliers* al estimar las varianzas y covarianzas de los efectos aleatorios del modelo directamente de los residuos. También se pueden emplear los métodos de remuestreo (*Bootstrapping*), que estiman los parámetros de un modelo y sus errores estándar de la muestra en una distribución muestral empírica.

Se pueden combinar métodos tradicionales con otros más complejos para la detección de *outliers.* Frumosu

& Kulahci (2019) propusieron una estrategia iterativa en la cual combinan la T^2 de Hotelling y el estadístico Q para emplearlos en una regresión de componentes principales semi-supervisada, en conjuntos de datos simulados y reales.

CONSIDERACIONES FINALES

En las investigaciones se pueden obtener, con frecuencia, observaciones atípicas que parecen no formar parte de esa muestra o población estadística. Su presencia puede ser problemática en el análisis de los datos, porque conlleva, al incumplimiento de las premisas de los métodos paramétricos, a sesgos en la estimación de los parámetros, a una disminución de la potencia de la prueba estadística y a incrementos en la tasa de error experimental, por lo que es de vital importancia determinar por qué un punto particular es un *outlier* en términos de su comportamiento relativo con respecto a los datos remanentes. Estos valores pueden ser resultado de fenómenos interesantes obtenidos por azar que vale la pena investigar con profundidad.

A pesar de que se han desarrollado un gran número de técnicas paramétricas y no paramétricas para la identificación y manejo de los *outliers,* sigue faltando en la literatura científica relacionada con el tema, una guía o procedimiento que tenga en cuenta la forma en que deben ser identificados y tratados estos valores, en diversos contextos experimentales.

Es importante que los investigadores sean cautelosos durante el muestreo, la conducción del experimento y la entrada de los datos al programa estadístico, para evitar los *outliers* por errores que pueden ser corregidos. También, debe existir correspondencia entre la definición que manejan del concepto y el método que seleccionan para identificarlos y manejarlos, y que tengan en cuenta si los *outliers* están presentes para una o múltiples variables.

Finalmente, se debe tener en cuenta que existen algunos estimadores y métodos cuyos resultados pueden estar influenciados por la presencia de los valores atípicos, por lo que se recomienda el empleo de estimadores y métodos robustos para el análisis de los datos, los cuales están disponibles en diferentes programas estadísticos.

AGRADECIMIENTOS

La autora desea agradecer a María Teresa Cornide y Majela Hernández, así como a los revisores anónimos y editores de la Revista del Jardín Botánico Nacional por sus revisiones críticas y constructivas del manuscrito.

REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C.C. 2013. Outlier Analysis. Springer, IBM T.J. Watson Research Center, Yorktown Heights. New York, USA.

Aguinis, H., Gottfredson, R.K. & Joo, H. 2013. Best-practice recommendations for defining, identifying and handling outliers. *Organ. Res. Methods* 16(2): 270-301.

Angiulli, F., Basta, S. & Pizzuti, C. 2006. Distance-based detection and prediction of outliers. *IEEE T. Knowl. Data En.* 18: 145-160.

Barnett, V. & Lewis, T. 1994. Outliers in Statistical Data. 3er Ed. John Wiley & Sons. New York, USA.

Breunig, M.M., Kriegel, H.P., Ng, R.T. & Sander, J. 2000. LOF: identifying density-based local outliers. *SIGMOD Rec.* 29(2): 93-104.

Cleophas, T.J. & Zwinderman, A.H. 2019. Outliers assessed as dependent adverse effects. En: Analysis of safety data of drug trials: An Update. Springer Nature Switzerland AG. Cham, Switzerland.

Dan, E. & Ijeoma, O.A. 2013. Statistical analysis/methods of detecting outliers in a univariate data in a regression analysis model. *International Journal of Education and Research* 1(5): 302-337.

De Armas, A.A. 2015. Detección de outliers en grandes bases de datos. Tesis de Maestría. Universidad Argentina de la Empresa, Argentina.

Evans, V.P. 1999. Strategies for detecting outliers in regression analysis: An introductory primer. En: Advances in Social Science Methodology. B. Thompson (Ed.). JAI Press, Stamford, Connecticut, USA.

Finch, W.H. 2012. Distribution of variables by method of outlier detection. *Front. Psychology* 3: 211.

Frumosu, F.F.& Kulahci, M. 2019. Outliers detection using an iterative strategy for semi-supervised learning. *Qual Reliab Engng Int.* 1-16.

García, Ch. 2017. How to Find Outliers in a Data Set. Academy Resources. www.http://blog.socialcops.com/academy/resources/find-deal-outliers-data-set/. 10 de septiembre de 2019.

Garson, G.D. 2012. Testing Statistical Assumptions. G.D. Garson and Statistical Associates Publishing. Asheboro, North Carolina, USA.

Hawkins D.M. 1980. Identification of outliers. Champman & Hall. London, UK.

Herrera, M., Guerra, C.W., Sarduy, L., García, Y. & Martínez, C.E. 2012. Diferentes métodos estadísticos para el análiis de variables discretas. Una aplicación en las ciencias agrícolas y técnicas. *Rev. Cie. Tec. Agr.* 21(1): 58-62.

Iglewicz, B. & Hoaglin, D. 1993. How to detect and handle outliers. ASQC Quality Press. Milwaukee, Wisconsin, USA.

Johnson, R.A. & Wichern, D.W. 1992. Applied Multivariate Statistical Analysis. 3rd Ed. Prentice Hall, Englewood Cliffs. New Jersey, USA.

Judd, C.M., McClelland, C.H. & Ryan, C.S. 2009. Data analysis: a model-comparison approach. 2nd Ed. Routledge. New York, USA.

Knorr, E. M. & Ng, R.T. 1998. Algorithms for mining distance-based outliers in large datasets. Pp. 392-403. En: Proceedings of the

24th International Conference on Very Large Data Bases. New York, USA.

Kulich, C., Trojanowski, G., Ryan, M.K., Haslam, S.A. & Renneboog, L.D.R. 2011. Who gets the carrot and who gets the sick? Evidence of gender disparities in executive remuneration. *Strategic Manage. J.* 32: 301-321.

Kwak, S.K. & Kim, J.H. 2017. Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology* 70(4): 407-411.

Muñoz, J.A. & Amón, I. 2013. Técnicas para detección de outliers multivariantes. *Revista en Telecomunicaciones e Informática* 3(5): 11-25.

Murphy, T. & Lau, A.T. 2008. Manejo de valores atípicos. ¿Cómo se evalúa un valor aberrante o inconsistente único? ASTM Standarization News.

Osborne, J.W. & Overbay, A. 2004. The power of outliers (and why researches should always check for them). *Pract. Assess. Res. Eval.* 9(6): 1-8.

Osborne, J.W. 2002. Notes on the use of data transformations. *Pract. Assess. Res. Eval.* 8(6): 1-9.

Osborne, J.W., Christiansen, W.R.I. & Gunter, J.S. 2001. Educational Psychology from a statistician's perspective: A review of the quantitative quality of our field. En: Proceedings of the Annual Meeting of the American Educational Research Association. Seattle, Washington, USA.

Pamula, R., Deka, J.K. & Nandi, S. 2011. An Outlier Detection Method based on Clustering. Second International Conference on Emerging Applications of Information Technology. Pp. 253-256. IEEE Computer Socierty, Kolkata, India.

Parrinello, C.M., Grams, M.E., Sang, Y., Couper, D., Wruck, L.M., Li, D., Eckfeldt, J.H., Selvin, E. & Coresh, J. 2016. Iterative Outlier Removal: A Method for Identifying Outliers in Laboratory Recalibration Studies. *Clin. Chem.* 62(7): 966-972.

Pérez, L. 2018. ¿Cómo proceder ante el incumplimiento de las premisas de los métodos paramétricos? o ¿Cómo trabajar con variables biológicas no normales? *Revista Jard. Bot. Nac. Univ. Habana* 39: 1-12.

Ramaswamy, S., Rastogi, R. & Shim, K. 2000. Efficient algorithms for mining outliers form large data sets. Pp. 427-438. En: Proceedings of International Conference on Management of Data, Dallas, Texas, USA.

Rousseeuw, P.J. & Leroy, A.M. 2003. Robust Regression and Outlier Detection. John Wiley & Sons. New York, USA.

Rousseeuw, P.J. & van Driessen, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41: 212-223.

Sandbhor, S. & Chaphalkar, N.B. 2019. Impact of outlier on neural networks based property value prediction. *Advances in Intelligent systems and Computing* 862: 481-495.

Seo, S. 2006. A review and comparison of methods for detecting outliers in univariate data sets. Tesis de Maestría. University of Pittsburg, USA.

Sykacek, P. 1997. Equivalent Error Bars for Neural Network Classifiers Trained By Bayesian Inference. Pp. 121-126. En: Proceedings of the European Symposium on Artificial Neural Networks. Bruges, Belgium.

Van der Loo, M.P.J. 2010. Distribution based outlier detection in univariate data. Statistics Netherlands. The Hague/Heerlen, Netherlands.

Whitlock, M.C. & Schluter, D. 2009. The Analysis of Biological Data. Roberts and Company Publishers. Grenndwood Village, Colorado, USA.

Wilcox, R.R. 2005. Introduction to Robust Estimation and Hypothesis Testing. Elsevier Academic Press. Burlington, Massachussets, USA.

Zar, J.H. 2010. Biostatistical Analysis. 5th Ed. Pearson Prentice Hall. New Jersey, USA.

Zhang, J. 2013. Advancements of Outlier Detection: A Survey. *ICST Transactions on Scalable Information Systems* 13(01-03): e2.

Zhang, K., Hutter, M. & Jin, H. 2009. A new local distance-based outlier detection approach for scattered real-world data. En: Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand. 813-822.