APRENDE CON ELI

ESTADÍSTICA

Análisis Robusto de Datos

TABLA DE CONTENIDO

¿QUÉ ES UN ATÍPICO?

Los **datos atípicos o outliers**, o también conocidos como datos anómalos o anomalías, son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Desde el punto de vista distribucional, se podría decir que son datos que no pertenecen a la misma distribución que los demás, es decir, han sido creados por un proceso o una distribución diferente.

Los datos atípicos pueden ser ocasionados por diversas causas:

a) Errores de procedimiento o de medición.

b) Acontecimientos extraordinarios.

c) Valores muy extremos.

d) Causas no conocidas.

EJEMPLOS

Por ejemplo, una muestra de datos del número de cigarrillos consumidos a diario contiene el valor 60 porque hay un fumador que fuma sesenta cigarrillos al día. Los demás datos son fumadores que no fuman más de 10 cigarrillos, entonces ese fumador es atípico en el sentido de que es un extremo y está muy alejado de lo que sería un fumador estándar.

Otro ejemplo puede ser las medidas de altura de 100 personas, se recogen en cm, sin embargo, por un error de medición las medidas de las últimas 10 personas se han recogido en metros. Esos valores son atípicos por el error de medición porque sus alturas van a ser números muy pequeños (por ejemplo 1.7 metros) comparados con las alturas de las 90 personas que se les midió en cm (por ejemplo 168 cm). Así que esas 10 personas forman un grupo de outliers en la muestra y pueden distorsionar las estimaciones y los análisis que hagamos sobre nuestros datos.

CONSECUENCIAS

Los datos atípicos distorsionan los resultados de los análisis, y por esta razón hay que identificarlas y tratarlos de manera adecuada. Generalmente excluyéndolos del análisis, pero en ocasiones no es necesario.

Si nos preguntan qué medida describe mejor un conjunto de datos, ¿qué responderías? La mayoría de las personas diría rápidamente… ¡la media!, pero veamos si es la respuesta correcta.

Supongamos que tenemos los siguientes datos:

10,10,11,12,12,13,14,15,15,15,16,18,19

Calculamos su media, que es 13.85.

¿Qué pasa si cambiamos un dato? Cambiemos el último número, el 19 y pongamos que es 200:

10,10,11,12,12,13,14,15,15,15,16,18,200

Volvemos a calcular la media, que ahora es 27.77, bastante superior a la anterior que era 13.85.

Con mover **solamente un valor** lejos del resto, ¡la media lo seguirá! Hemos puesto un valor más grande que el resto, y la media ha aumentado muchísimo. ¡Un solo valor es suficiente para influenciar enormemente la media del conjunto de datos!

Este ejemplo indica que la **robustez** del estimador es importante cuando tenemos datos atípicos (u outliers) y también cuando no queremos que un dato tenga más influencia que los demás en los cálculos.

SOLUCIONES

En este caso, hay dos soluciones. Una sería **eliminar** el dato atípico, en este caso es solo una observación, pero en otros casos podemos encontrarnos con que son muchas más de las que nos gustarían y no queremos perder información.

Si hemos corroborado que estos valores atípicos no se deben a un error a la hora de construir la base de datos o en la medición de la variable, eliminarlos puede ser que no sea la mejor solución.

Si no se debe a un error, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esa información, debido a que introduce un sesgo, a que disminuye el tamaño muestral y a que puede afectar tanto a la distribución como a las varianzas.

Además, ¡en la variabilidad de los datos reside el tesoro de nuestra investigación!

Es decir, la variabilidad (diferencias en el comportamiento de un fenómeno) debe explicarse no eliminarse. Y si aún no puedes explicarla al menos debes poder disminuir la influencia de estos valores atípicos en tus datos.

Podemos interpretar que los datos atípicos «pesan más» que los datos cercanos a la media. Entonces la mejor opción es quitarles **peso** a esas observaciones atípicas.

Esas dos van a ser las ideas básicas de las **técnicas robustas**.

Los métodos estadísticos robustos son técnicas modernas que hacen frente a estos problemas. Son similares a los clásicos, pero se ven menos afectados por la presencia de valores atípicos o variaciones pequeñas respecto a las hipótesis de los modelos.

ESPACIO UNIVARIANTE

El espacio univariante es el caso en el que tenemos una sola variable aleatoria, es decir, medimos solo una característica para todos los datos que tenemos.

Este caso es el más sencillo, vamos a ver cómo pueden afectar los atípicos a muchas de las estimaciones que estamos acostumbrados a hacer con nuestros datos.

Un solo valor atípico puede tener un efecto enorme sobre la **media**, como ya vimos anteriormente.

Vamos a considerar el siguiente **ejemplo** de 4 conjuntos de datos:

Conjunto de datos 1: 150, 160, 130, 150, 120

Conjunto de datos 2: 150, 160, 130, 150, 120, 180

Conjunto de datos 3: 150, 160, 130, 150, 120, 350

Conjunto de datos 4: 150, 160, 130, 150, 120, 300, 320, 340, 350.

El primero no tiene ningún valor atípico, el segundo tiene sólo un valor atípico relativamente moderado, el tercero tiene un valor atípico muy extremo y el último tiene varios valores atípicos muy extremos.

Si calculamos la media para cada conjunto de datos (CD):

La media de CD1 es 142.

La media de CD2 es 148,3.

La media de CD3 es 176.

La media de CD4 es 224.

Vemos que la media cambia bastante entre los conjuntos de datos, pero notemos cómo es mucho más alta mientras más extremos son los valores de los atípicos.

Vamos a calcular ahora la mediana.

La mediana para los conjuntos de datos 1, 2 y 3 es 150. Y en el conjunto de datos 4, solo aumenta hasta 160. La mediana se ve mucho menos afectada por los valores atípicos que la media. Esto es porque la mediana es un estimador robusto, no se deja influenciar demasiado por la presencia de valores atípicos.

Esto demuestra que debemos tener cuidado con las estimaciones que hacemos, porque por supuesto, cualquier otro estimador que dependa de la media va a estar en consecuencia también afectado y cualquier método que esté basado en ese estimador le pasará lo mismo. Por ello, muchos procedimientos que están basados en la media tienen una versión alternativa que usa la mediana en su lugar para obtener un método robusto.

Otro caso es la desviación típica, que podemos interpretar como un promedio de las desviaciones con respecto a la media. Por lo cual también se verá afectado. Una alternativa es el MAD, que en inglés es: *median absolute deviation from the data's median*, es decir, la mediana de la desviación absoluta con respecto a la mediana. Como ves es la misma idea que la desviación típica, pero usando la mediana en vez de la media.

Un resumen de estimadores no robustos vs alternativas robustas es el siguiente:

Estimadores no robustos	Alternativas robustas
Media	Mediana
Desviación típica	MAD
Asimetría	Medcouple

A la hora de identificar a los atípicos hay dos problemas muy importantes:

Masking effect:

El masking effect o efecto de **enmascaramiento** ocurre hay un valor atípico o varios muy extremos que dejan enmascarados a otros atípicos intermedios y lo que sucede es que el método no los detecta como valores atípicos.

Swamping effect:

Por otro lado, el swamping effect o efecto de pantano o **empantanamiento**, puede ocurrir cuando hay tantos valores atípicos extremos que sucede que algunas observaciones que en realidad no son atípicas al final son declaradas como outliers, porque las otras eran demasiado influyentes o extremas y han trastornado el método, haciendo que detecte falsos positivos.

Algunos métodos para detectar atípicos que NO son robustos y pueden sufrir de estos efectos, junto con sus alternativas robustas son:

Métodos no robustos	Alternativas robustas
Método SD	MADe
Método Zscore	Zsocre modificado
Boxplot	Boxplot ajustado

ESPACIO MULTIVARIANTE

En el espacio multivariante, cuando tenemos más de una variable aleatoria, los análisis se complican. Porque la definición de atípico ya no está tan clara. ¿Sería una observación que tiene valores extremos en todas las variables que tenemos? ¿O podría ser extremo sólo en una de esas variables? Entonces lo mejor es recurrir al concepto distribucional, medir de alguna manera si

distribucionalmente hay observaciones alejadas del grueso mayoritario de datos que tenemos. Para ello necesitamos el concepto de distancia, ya que tampoco es fácil ordenar los datos en el caso multivariante, el orden no es igual que cuando sólo tenemos una variable aleatoria que al final es una lista de datos cuyos valores podemos ordenar fácilmente.

En este caso, usaremos mayormente la distancia de Mahalanobis que mide cuán alejados están los datos del centro teniendo en cuenta su dispersión. El problema de la definición clásica de la distancia de Mahalanobis es que depende de dos estimadores no robustos: el vector de medias y la matriz de covarianza. Estos dos son los equivalentes multivariantes de la media y la desviación típica y la varianza que teníamos en el caso univariante. Además, en multivariante también tendremos otras dos medidas que son la covarianza y la correlación que miden esa interrelación que hay entre dos variables. Todas ellas: media, desviación típica, varianza, covarianza y correlación sufren con la presencia de atípicos y pueden dar estimaciones puntuales erróneas si no lidiamos con ellos. Por eso si se utilizan en la distancia de Mahalanobis para detectarlos, realmente el método puede no ser eficaz si está basado en estimaciones erróneas del centro y la dispersión de los datos. Y terminaremos no detectando a todos los datos problemáticos, algunos estarán escondidos (masking) o incluso puede que otros terminen siendo detectados como atípicos sin realmente serlo (swamping).

Por eso surge el concepto de distancia de Mahalanobis robusta, que es simplemente la misma idea de antes, en vez de usar los estimadores sensibles, usemos estimadores robustos a atípicos, así la distancia tendrá la misma propiedad y tendremos menor probabilidad de equivocarnos.

¿Cuáles son las alternativas robustas a esos estimadores?

No robusto	Robusto
Vector de medias	Mediana geométrica o L1
Matriz de covarianza	Matriz comedian
Matriz de correlaciones	Matriz correlation-median

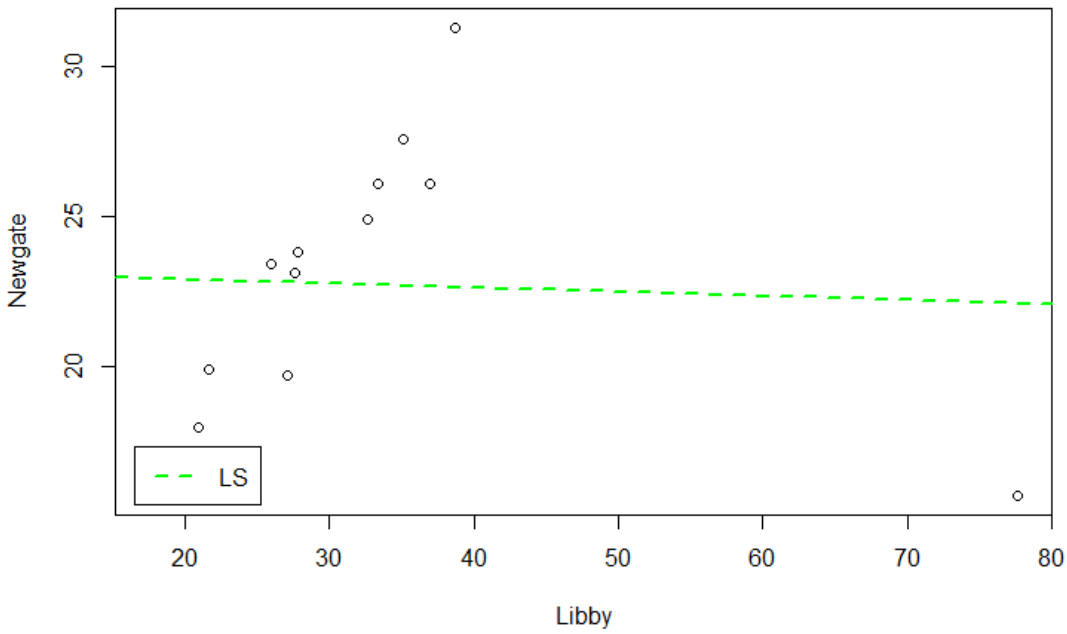
Otras alternativas son métodos que en un solo procedimiento (casi siempre iterativo) estiman el centro de los datos (también conocido como localización) y la dispersión, es decir, la matriz de covarianza, de manera robusta, ya sea eliminando atípicos en cada caso o dándole menos peso para hacer las estimaciones en cada iteración, para al final devolver dos estimaciones finales que sean robustas.

- MCD
- MCD ajustado
- Stahel-Donoho
- Método de la kurtosis

REGRESIÓN LINEAL

La regresión lineal es un caso interesante, donde se puede interpretar que los datos anómalos son aquellos que no siguen el patrón de regresión lineal que siguen los demás.

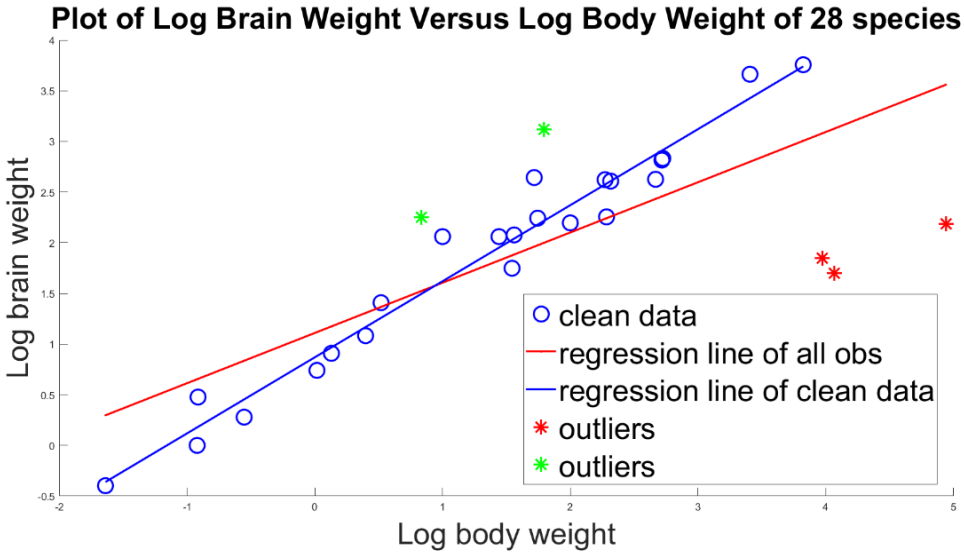
Un ejemplo muy ilustrativo es el del conjunto de datos llamado Kootenay, es clásico en los libros de regresión robusta. Aquí se mide el caudal del rio Kootenay entre 1971 y 1973 en dos lugares, uno llamado Libby y otro llamado Newgate. La idea es predecir el caudal en Newgate una vez que se observa lo que pasa en Libby. Cuando se ajusta la línea de regresión según el método clásico resulta lo siguiente:



Como se puede ver la gran mayoría de datos está a la izquierda y sigue un patrón de regresión lineal positivo, sin embargo, la recta (verde) de regresión estimada es completamente errónea, no representa ese patrón e incluso tiene pendiente negativa lo cual es inverso al verdadero patrón. ¿Cuál es la causa? Que en la esquina inferior derecha hay un dato atípico muy influyente que trastorna por completo la estimación de los parámetros del modelo de regresión.

Otro ejemplo muy común es el caso del dataset de 28 especies animales, mamíferos específicamente, donde se miden dos características de cada uno de esos animales que son el peso de su cuerpo y el peso de su cerebro. En este ejemplo lo que se observa es que existe para la mayoría de los datos una relación lineal positiva entre ambas medidas, es decir, que mientras más pesa el cuerpo del animal más pesa su cerebro también. Sin embargo, hay 5 observaciones, es decir, 5 especies que no cumplen con ese patrón, sino que de alguna manera difieren y están alejados de lo que sería la regresión que se obtendría para los demás datos. Entonces ocurre algo muy interesante, si se estima la recta de regresión para todos los datos el resultado es muy diferente que si por ejemplo se eliminan esos 5 datos que difieren de los

demás y se estima la recta de regresión para los datos restantes, como se ve en la siguiente imagen.



Los datos en color verde y rojo son los atípicos, la recta de color azul sería la regresión estimada solo con los 23 datos restantes después de eliminar los 5 atípicos, y la recta de color rojo es el resultado de considerar todos los 28 datos. Vemos cómo cambia el resultado. Esto es porque esas observaciones atípicas influyen mucho en las estimaciones de los parámetros de la recta de regresión haciendo que el resultado final sea incorrecto si se incluye a los outliers en la muestra al estimar la recta.

Sin embargo, en la práctica la idea de eliminar las observaciones puede no ser del todo buena. Es por eso que en regresión existen métodos de regresión robusta que lo que hacen básicamente es darles menor peso a esas observaciones, identificándolas casi siempre en base a cuán influyentes son respecto al comportamiento de los residuos (errores) del modelo.

El método clásico que sabemos que no es robusto es:

- LS: Least squares – mínimos cuadrados

Los métodos robustos más utilizados son:

- LAD: Least absolute deviation – desviación minima absoluta
- LMS: Least median deviation – desviación mínima de la mediana
- LTS: Least trimmed squares – mínimos cuadrados recortados

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA) ROBUSTO

El método clásico de Análisis de componentes principales (PCA) presenta problemas cuando hay presencia de atípicos en los datos.

Un breve recordatorio: PCA es un método para reducir la dimensión de los datos, donde las nuevas variables van a ser combinaciones lineales de las variables originales. Por ejemplo, si tenemos un gran número de variables en nuestro dataset y además con alta dependencia, PCA va a ser muy útil a la hora de reducir la dimensión, porque sabemos que si tenemos variables multicorrelacionadas eso significa que habrá redundancia de información y que un grupo de menor tamaño puede aportar prácticamente la misma información y a la vez se disminuye bastante la complejidad al considerar menos variables, porque se eliminan las variables redundantes.

Entonces con PCA, antes de hacer otro tipo de análisis como regresión, clasificación, etc, podemos deshacernos de algunas de esas variables y así reducir la dimensión. Que no es nada más que convertir nuestros datos en un dataset más sencillo de analizar y más sencillo de interpretar. Por lo tanto, este método de reducción de la dimensión, si tenemos variables de alta dependencia, es útil porque luego nos va a dar un número pequeño de nuevas variables que van a explicar la mayor parte de la variabilidad original, así que no se va a perder demasiada información. Entonces, el resumen es que se transforman las variables originales, que en general van a estar correadas, en nuevas variables que están incorreladas, por lo tanto, esto es una transformación ortogonal, lo cual facilita la interpretación de los datos. Además, los componentes principales se calculan como la combinación lineal de las variables originales que tenga varianza máxima.

Si quieres saber más detalles del método clásico de componentes principales (PCA), lo tenemos como una sección completa en el *Curso avanzado estadística multivariante* que lo pueden encontrar mi en mi perfil de Udemy o en aprendeconeli.com donde además vendrá con descuento.

El método PCA clásico, para hallar los componentes, se basa en el estimador clásico de la matriz de varianzas y covarianzas. Por lo tanto, esto va a tener un problema cuando tenemos atípicos, pues este estimador es sensible a su presencia y puede estar gravemente influenciado por ellos, y al final terminaremos con una transformación equivocada.

Además, la presencia de atípicos sabemos que es muy difícil de detectar en grandes dimensiones. Entonces este método es muy útil para, primero, detectar grupos de atípicos y reducir la dimensión, reduciendo así también la complejidad de los datos. Pero no podemos utilizarlo si nos basamos en el estimador clásico de la varianza. Por lo tanto, el desafío es encontrar métodos confiables, que también sean rápidos, robustos ante valores atípicos y aplicables a datos de alta dimensión. Por esto, sobre todo por la parte de alta dimensión, vamos a mencionar también otro método llamado Sparse PCA que también tiene en su versión clásica su versión robusta.

Sparse significa disperso, esparcido o poco denso. Nosotros nos vamos a referir a sparse, que es como se dice en la literatura científica al referise a este método, aunque esté en inglés.

SPARSE PCA

Cuando los datos son sparse, lo que pasa es que los puntos van a estar dispersos, casi equidistantes entre sí. En otras palabras, el uso de distancias euclídeas, por ejemplo, no va a tener sentido. Entonces el resultado es que el grado de periferia de los puntos de datos va a ser indistinguible y, por tanto, los valores atípicos se van a detectar mejor mediante el uso de un subespacio local de menor dimensión, en el que solo un subconjunto de características es relevante.

Existe una amplia gama de enfoques disponibles, por ejemplo, dentro del análisis de componentes principales, que es una técnica especializada dentro del análisis estadístico y en particular del análisis de datos multivariante. Es decir, existen extensiones del método clásico PCA que por un lado lidia con el problema atípicos y por otro lado con el de la alta dimensionalidad.

La única desventaja del PCA clásico no es que no sea robusto sino también que los componentes principales suelen ser combinaciones lineales de todas las variables originales. Por lo tanto, esto puede ser una desventaja a la hora de la interpretación porque estoy considerando una combinación lineal donde uso todas las variables y no necesariamente tiene que pasar eso, puede ser que algunas no sean relevantes ni siquiera para los componentes. Por tanto, la versión de Sparse PCA tiene la ventaja de encontrar combinaciones lineales que contienen solo unas pocas variables de entrada.

Si por ejemplo sucede el coeficiente para una variable en el primer componente principal es cero, no estoy considerando esa variable para ese componente, y esto facilita la interpretación.

¿QUÉ APLICACIONES TIENE ESTO?

Por ejemplo, se usa mucho en análisis de datos financieros. Supongamos que el PCA clásico se aplica a un conjunto de datos donde cada variable de entrada representa un activo diferente. Entonces podemos generar componentes principales que son una combinación ponderada de todos los activos. Por el contrario, el Sparse PCA produciría componentes principales que son una combinación ponderada de solamente unos pocos activos. Por lo que uno puede interpretar fácilmente su significado. Además, si se utiliza

una estrategia comercial basada en estos componentes principales, menos activos implican menos costo de transacción.

Otro ejemplo, en el campo de la biología, específicamente en genética, puede ser si consideramos un conjunto de datos donde cada variable de entrada, cada variable original corresponde a un gen específico, pues el Sparse PCA puede producir un componente principal que involucre sólo unos pocos genes, por lo que los investigadores pueden enfocarse en estos genes específicos para su posterior análisis.

Decía C.C. Aggarwal que los valores atípicos a menudo están ocultos en un comportamiento local inusual de los subespacios de baja dimensión, y este comportamiento desviado está enmascarado por el análisis en altas dimensiones.

LIBRERÍAS EN R

Se han desarrollado varios enfoques para abordar de manera robusta el PCA. Por ejemplo, hay dos librerías en R, que son el pcaPP y el rospca. El primero es un PCA robusto basado en la búsqueda de proyecciones. Es un enfoque introducido inicialmente en el año 1985. Lo que pasa es que ese algoritmo era muy lento en la práctica, así que luego fueron surgiendo modificaciones nuevas y mejores algoritmos para la misma idea. La idea es proyectar los datos en un espacio de dimensión inferior, tal que se maximice una medida robusta de la varianza de los datos proyectados, es decir, considerar un estimador robusto en vez del estimador clásico, que es el que se utiliza en PCA.

Por otro lado, rospca es un paquete que contiene dos métodos de PCA robusto. Uno viene en la función robpca, que combina el enfoque de búsqueda proyecciones con la estimación robusta de la covarianza basada en un estimador de la matriz de confianza que hemos visto en clase, que es el MCD. Es bastante más rápido que otras alternativas anteriores. Además, tienen la ventaja de ser aplicable tanto a datos distribuidos simétricamente como datos asimétricos. Y también está la versión sparse de este PCA robusto de la mano de la función rospca dentro del paquete.

CONCLUSIONES

En resumen, podemos tener datos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no

se puede, deberían ser identificados y eliminarse del análisis o recodificarse como datos ausentes. También podemos tener observaciones que ocurren como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento valido de la población y puede ser eliminado del análisis. Otro caso serían las observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis, pero estudiando que influencia ejercen en los procesos de estimación de los modelos considerados. Y finalmente, pueden existir datos extraordinarios para los que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el porqué de dichas observaciones.

Las técnicas robustas del análisis de datos son aquellas que se encargan de, ya sea detectar los atípicos para eliminarlos del análisis o identificarlos para ponerle menos peso a esas observaciones sin tener que eliminarlas del todo. Lo más interesante de esta idea, es que las técnicas clásicas para detectar atípicos como el boxplot en el caso univariante, o la distancia de Mahalanobis clásica en el caso multivariante, no son técnicas robustas, por lo cual van a estar influenciadas por los atípicos y van a sufrir dos problemas claves que son el “masking” y el “swamping”. Es decir, que los propios atípicos pueden afectar el método para su detección y el procedimiento falla en identificarlos a todos. Por eso el estudio de técnicas que son completamente robustas en el sentido de que todos los estimadores que se usan en los pasos del método no se dejan influenciar por los atípicos.