Valores extremos o atípicos en una muestra estadística

¿La liga española de futbol de la primera división está “dopada”?

Apellidos, nombre	Boigues Planes, Francisco José[1] (fraboipl@mat.upv.es) Estruch Fuster, Vicente Domingo[2] (vdestruc@mat.upv.es)
Departamento	[1,2]Departamento de Matemática Aplicada
Centro	Universitat Politècnica de València

1 Resumen de las ideas clave

Presentamos un enfoque al tratamiento de los valores extremos o atípicos (outliers en inglés). En concreto, se pretende dar una noción de valor "atípico" en una muestra estadística y además se describirá la forma de hallarlos. Con ello reforzaremos las ideas sobre la representatividad de las medidas de tendencia central y de dispersión en los estudios de estadística descriptiva o en los análisis exploratorios de datos.

2 Introducción

Las medidas de tendencia central de una muestra (media, mediana, moda, etc.) pretenden resumir en un único valor dicho conjunto de datos. La medida de tendencia central más utilizada es la media aritmética, pero si pretendemos representar todo el conjunto de datos mediante la media aritmética hemos de ser muy precisos. Si nuestros datos presentan algunos valores extremadamente altos o extremadamente bajos, en comparación con el resto, la media aritmética también va a ser demasiado alta o demasiado baja, respectivamente, como para poder representar adecuadamente el conjunto de datos. Esta situación se podría evitar calculando la mediana y la moda, y comparando las tres medidas. Si las tres medidas tienen, más o menos, un valor similar, entonces se puede confiar en que tanto la media, la mediana como la moda, proporcionan un valor representativo.

Las medidas de dispersión, en su versión más general, miden el grado de dispersión de los valores de la variable respecto de la medida de tendencia central elegida. En este grupo tendríamos las desviaciones cuadráticas medias, las desviaciones típicas, o el coeficiente de variación. Otro grupo de medidas de dispersión sería aquel que pretende dar una idea de hasta qué punto se extienden los datos en la recta real. En este grupo tendríamos el rango, los cuartiles y el rango intercuartílico.

Un valor extremo o atípico en un conjunto de datos es un valor que es muy diferente del resto de valores. En la mayoría de los casos, los atípicos tienen influencia en la media, pero no en la mediana, o la moda. Es importante tener en cuenta que, en el caso del cálculo de la media aritmética, los datos atípicos son más influyentes que los datos cercanos a la media. Por otra parte, los atípicos también son importantes por su efecto en cualquiera de las medidas de dispersión.

En el desarrollo del artículo comenzaremos planteando una cuestión recurrente al comienzo de la liga española de futbol que nos llevara a una aproximación de la noción de valor atípico. A continuación, veremos un método para la obtención de datos atípicos conocido como método del diagrama de Box-Whisker. Finalmente, plantearemos varios problemas que refuercen lo aprendido.

3 Objetivos

Al finalizar este artículo, un estudiante debe ser capaz de:

- Caracterizar la noción de "valor atípico" de una muestra estadística.
- Utilizar técnicas para determinar los valores atípicos de una muestra.

4 Desarrollo

Se inicia la sesión con el planteamiento de la siguiente pregunta:

- **¿La liga española de futbol de la primera división está "dopada"?**

Como datos, se dispone del presupuesto para salarios (tope salarial) de los equipos de fútbol de la primera división de la liga española para la temporada 2018/2019 (en millones de euros) (Figura 1). Dichas cantidades corresponden al importe máximo en salarios que cada equipo puede consumir en la temporada 2018-2019. Las cantidades incluyen el gasto en el primer equipo en jugadores y en el cuerpo técnico.

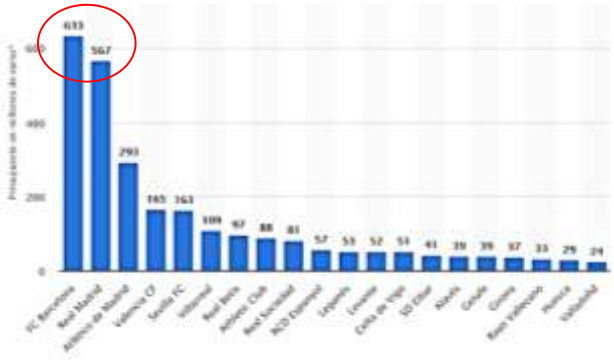

Figura 2.- Presupuesto de los equipos de fútbol de primera división, Liga BBVA 2018-2019 (Statista, 2018-2019)

Se puede acceder a dichos datos en Statista (2018-2019): https://es.statista.com/estadisticas/498947/presupuesto-equipos-de-futbol-de-la-liga-en-espana/

Traslademos los datos a una hoja Excel, de forma que en la casilla A1 escribiremos **Equipo**, e introduciremos los nombres de los equipos, sucesivamente, desde la casilla A2 a la casilla A21. En la casilla B1 escribiremos **Presupuesto**, y los valores correspondientes, sucesivamente, entre las casillas B2 y B21. De esta forma se tendría una tabla, con los valores de presupuesto por equipo, ordenados de mayor a menor (Figura 2)

Figura 2.- Tabla capturada en Excel con los presupuestos de los equipos

Completemos la información anterior con un simple análisis descriptivo. En primer lugar, calculamos la media aritmética, la mediana y la moda de todos los datos (incluyendo los atípicos). Para ello, utilizamos en Excel las funciones PROMEDIO, para la media aritmética; MEDIANA, para la mediana; y MODA.UNO para la moda (Tabla 1).

Tabla 1.- Parámetros de centralización de todos los equipos

MEDIDA	FUNCIÓN EXCEL O FÓRMULA	RESULTADO
Media aritmética global ($\bar{x}_G$):	=PROMEDIO(B2:B21)	132,55
Mediana global (m_G):	=MEDIANA(B2:B21)	55
Moda global (mo_G):	=MODA.UNO(B2:B21)	39

Observemos que, si consideramos el presupuesto de todos los equipos, 15 de los 20 equipos (un 75%) tendrían un presupuesto por debajo de la media. Por otra parte, la mitad de los equipos tienen un presupuesto por debajo (y la otra mitad por encima) de 55 millones de euros. El valor de la moda es poco interesante puesto que se obtiene de la información de sólo 2 equipos que tienen el mismo presupuesto.

El rango es un parámetro de dispersión muy interesante ya que la diferencia entre el equipo con mayor presupuesto y el equipo con menor presupuesto es de 609 millones de euros, es decir, hay una diferencia notable entre equipos en cuanto al presupuesto. Además, calculemos también la desviación típica, σ_G, (DESVEST.P) y el coeficiente de variación ($CV_G = \frac{\sigma_G}{\bar{x}_G} \times 100$) de todos los datos (Tabla 2). Se observan valores de dispersión muy altos.

Tabla 2.- Parámetros de dispersión de todos los equipos

MEDIDA	FUNCIÓN EXCEL O FÓRMULA	RESULTADO
Rango global (R_G):	*=Maximo-Mínimo*	609
Desviación típica global (σ_G):	=DESVEST.P(B2:B21)	168,22
Coeficiente de variación global (CV_G):	$\frac{\sigma_G}{\bar{x}_G} \times 100$	126,91%

4.1 Una primera aproximación a la NOCIÓN de valor "atípico"

Observamos en la Figura 1 que los presupuestos del F.C. Barcelona y del R. Madrid (rodeados en rojo) están a bastante distancia del resto. También destaca, aunque en menor medida, el presupuesto del Atlético de Madrid. Parece obvio que los datos de presupuesto de F.C. Barcelona y R. Madrid son "candidatos" a ser atípicos. Por tanto, diremos que un valor es **atípico** (en inglés outlier) si es una observación que es numéricamente distante del resto de los datos. Puesto que se trata de presupuestos reales (no se trata de datos erróneos) por lo que no procede eliminarlos del conjunto, sino evaluar su influencia.

Calculamos ahora la media, la mediana, la desviación típica y el coeficiente de variación, parciales, es decir sin considerar los valores de F.C. Barcelona y R. Madrid (Tabla 3).

Tabla 3.- Parámetros estadísticos sin el R. Madrid ni FC Barcelona

MEDIDA	FUNCIÓN EXCEL O FÓRMULA	RESULTADO
Promedio parcial ($\bar{x}_P$):	=PROMEDIO(B4:B21)	80,61
Mediana parcial (m_P):	=MEDIANA(B4:B21)	52,50
Desviación Típica parcial (σ_P)	=DESVEST.P(B4:B21)	65,93
Coeficiente de variación parcial (CV_P):	$\frac{\sigma_P}{\bar{x}_P} \times 100$	81,78%

Observemos (Tablas 1 y 3) que la media aritmética ha disminuido notablemente, de 168.22 a 80.61, es decir un 39.18%. La mediana apenas ha variado y las medidas de dispersión, también han disminuido notablemente. Estamos ante evidencias de que los valores eliminados tenían una influencia notable ante la media y la desviación. Por tanto, estamos ante datos atípicos. Sin embargo, el valor promedio (80.61) todavía es sensiblemente superior a la mediana (52.50), lo cual indica asimetría positiva o a la derecha.

4.2 Criterio de Tukey o del diagrama de Box-Whisker para la detección de atípicos

Hemos iniciado el análisis con un análisis exploratorio de los datos, detectándose posibles atípicos a partir de la observación. En este apartado, vamos a aplicar el criterio de Tukey, también llamado del diagrama de Box-Wisker, para la detección de atípicos. Para ello, lo primero que hacemos es calcular el primer y el tercer cuartil, es decir los percentiles 0.25 ($P_{0.25}$) y 0.75 ($P_{0.75}$), respectivamente y el rango intercuartílico (Tabla 4).

Tabla 4.- Cuartiles de toda la muestra

MEDIDA	FUNCIÓN EXCEL O FÓRMULA	RESULTADO
Primer cuartil (percentil $P_{0.25}$):	=PERCENTIL.INC(B2:B21;0,25)	39
Tercer cuartil (percentil $P_{0.75}$):	=PERCENTIL.INC(B2:B21;0,75)	122,5
Rango intercuartílico (RI)	$P_{0.75} - P_{0.25}$	83,5

Calculamos el valor de vez y media del rango intercuartílico, es decir, $1.5 \times RI = 125.25$ y de 3 veces el rango intercuartílico, $3 \times RI = 250.5$. Formamos ahora una tabla con las distancias de los valores a los cuartiles más próximos. Por supuesto, para calcular las distancias, sólo hay que considerar los valores inferiores y superiores al primer cuartil y tercer cuartil, respectivamente (se han sombreados en amarillo y verde en la Tabla 5).

Las distancias se situarán en la columna C. En la parte superior de la tabla 5 se destacan en amarillo los equipos y presupuestos superiores a $P_{0.75} = 122.5$, y, en la tercera columna, se indica la distancia desde el valor al tercer cuartil. En la parte baja de la misma tabla, se remarcan en verde los equipos y presupuestos inferiores a $P_{0.25} = 39$ y se indica también la distancia al primer cuartil. Por ejemplo, para el F.C. Barcelona, en la casilla C1, escribimos =ABS(B1-122.5) y el

resultado será 510.5. Para el Girona, en la casilla C18, escribimos =ABS(B18-38), siendo el resultado 2. De esta forma se completa la Tabla 5.

Tabla 5.- Valores del presupuesto y distancias de los valores mayores y menores al cuartil más cercano

Club	Presupuesto	Distancia
F.C. Barcelona	633	510,5
R. Madrid	567	444,5
Atl. Madrid	293	170,5
Valencia C.F.	165	42,5
Sevilla F.C.	163	40,5
Villareal	109	
Real Betis	97	
Ath. C. Bilbao	88	
R. Sociedad	81	
R.C.D. Español	57	
Leganés	53	
Levante	52	
Celta	51	
SD Eibar	41	
Alavés	39	
Getafe	39	
Girona	37	2
Rayo Vallecano	33	6
Huesca	29	10
Valladolid	24	15

El criterio del diagrama de Box-Wisker indicaría que son atípicos aquellos presupuestos correspondientes a las distancias remarcadas en azul, que son las que superan el valor 125.25 (no hay atípicos en la parte inferior).

Por lo tanto, el criterio señala como atípicos los presupuestos del F.C. Barcelona, R. Madrid y Atl. Madrid. Más concretamente, los presupuestos de F.C Barcelona y R. Madrid son atípicos lejanos (la distancia supera el valor de $3 \times RI = 250.5$) y el del Atl. Madrid atípico leve o cercano. Con la información obtenida, podemos construir a mano el diagrama Box-whisker (Figura 3) donde los 3 atípicos, uno cercano y dos lejanos, quedan perfectamente identificados a la derecha del diagrama.

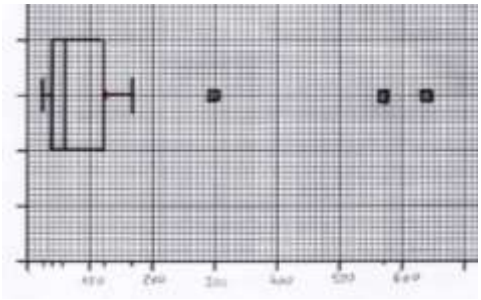

Figura 3.- Diagrama Box-whisker

4.3 El diagrama Box-whisker de Excel

Para obtener el diagrama de Box-whisker con Excel, actuamos de la siguiente manera:

- Seleccionamos las celdas de valores de **Presupuesto**.

- Seleccionamos *Insertar* e ir a *ver todos los gráficos* y elegir la opción correspondiente a *Cajas y Bigotes* (Figura 4):

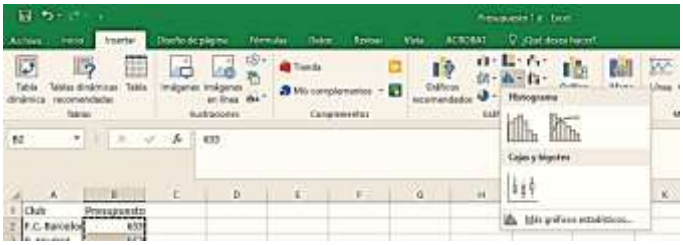

Figura 4.- Pasos para obtener el diagrama Box-whisker de Excel

- Obtenemos el gráfico de la Figura 5

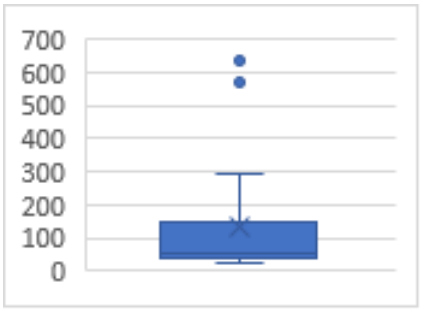

Figura 5.- Diagrama de Box-Whisker de Excel

- Observemos que, por defecto, Excel actúa de forma que en el diagrama (Figura 5) sólo se indican los atípicos lejanos (F.C. Barcelona y R. Madrid).

4.4- Respuesta a la pregunta planteada

En los presupuestos de los equipos de futbol de la liga BBVA existen grandes diferencias, por lo que la media aritmética no es una medida representativa para todos los equipos. El valor de la mediana indica que el 50% de los equipos tienen un presupuesto que no supera los 39 millones de €. La mitad de los equipos tienen un presupuesto entre 39 y 122 millones de € y más del 25% de los equipos no superan los 40 millones de euros. Los tres equipos con mayor presupuesto (F.C. Barcelona, R. Madrid y Atl. Madrid) acumulan un presupuesto (1493 mill. €) bastante mayor que el acumulado por el resto de los equipos (1158 mill. €). Consecuentemente, hay diferencias significativas en los presupuestos de los equipos de primera división.

4.5- Reforzando lo aprendido

Conviene consolidar lo aprendido mediante algún ejercicio de refuerzo como los propuestos a continuación:

Según la revista Chemical Engineering, una propiedad importante de una fibra es su absorción del agua. Se toma una muestra aleatoria de 20 pedazos de fibra de algodón y se mide la absorción de cada uno. Los valores de absorción son los siguientes:

18.71, 14.1, 20.72, 21.81, 19.29, 22.43, 20.17, 23.71, 19.44, 20.50, 18.92, 20.33, 23.00, 22.85, 19.25, 21.77, 22.11, 30.01, 18.04, 21.12

a) *Elabora una gráfica de puntos con los datos de la absorción y estima visualmente si existe algún posible atípico.*
b) *Calcula la media, la desviación típica y la mediana muestrales para los valores de la muestra anterior.*
c) *Calcula el primer y el tercer cuartil.*
d) *Elimina los posibles atípicos y calcula de nuevo las medidas del apartado b). Comenta los resultados.*
e) *Aplica el criterio del diagrama de Box-whisker para detectar los posibles atípicos.*
f) *Redacta un pequeño resumen de la muestra, bajo la hipótesis de que los posibles atípicos no provienen de ningún error.*

Otro problema interesante sería contrastar las noticias aparecidas en la prensa local (Figura 6) resaltando las temperaturas inusuales del mes de febrero en la comunidad valenciana.


LP Comunitat | El día de invierno más caluroso en Valencia en 30 años

Comunitat Valencia Alicante Castellón L'Horta - Morvedre La Marina La Safor

El día de invierno más caluroso en Valencia en 30 años

El termómetro llega a 27,3ºC en el aeropuerto y los registros baten el récord en invierno desde febrero de 1990

Figura 6.- Recorte de prensa (Fuente: Las Provincias, 27/02/2019)

Para ello disponemos de un gráfico de las temperaturas medias desde 1050. A partir del gráfico (Figura 7) se pueden estimar dichas temperaturas.

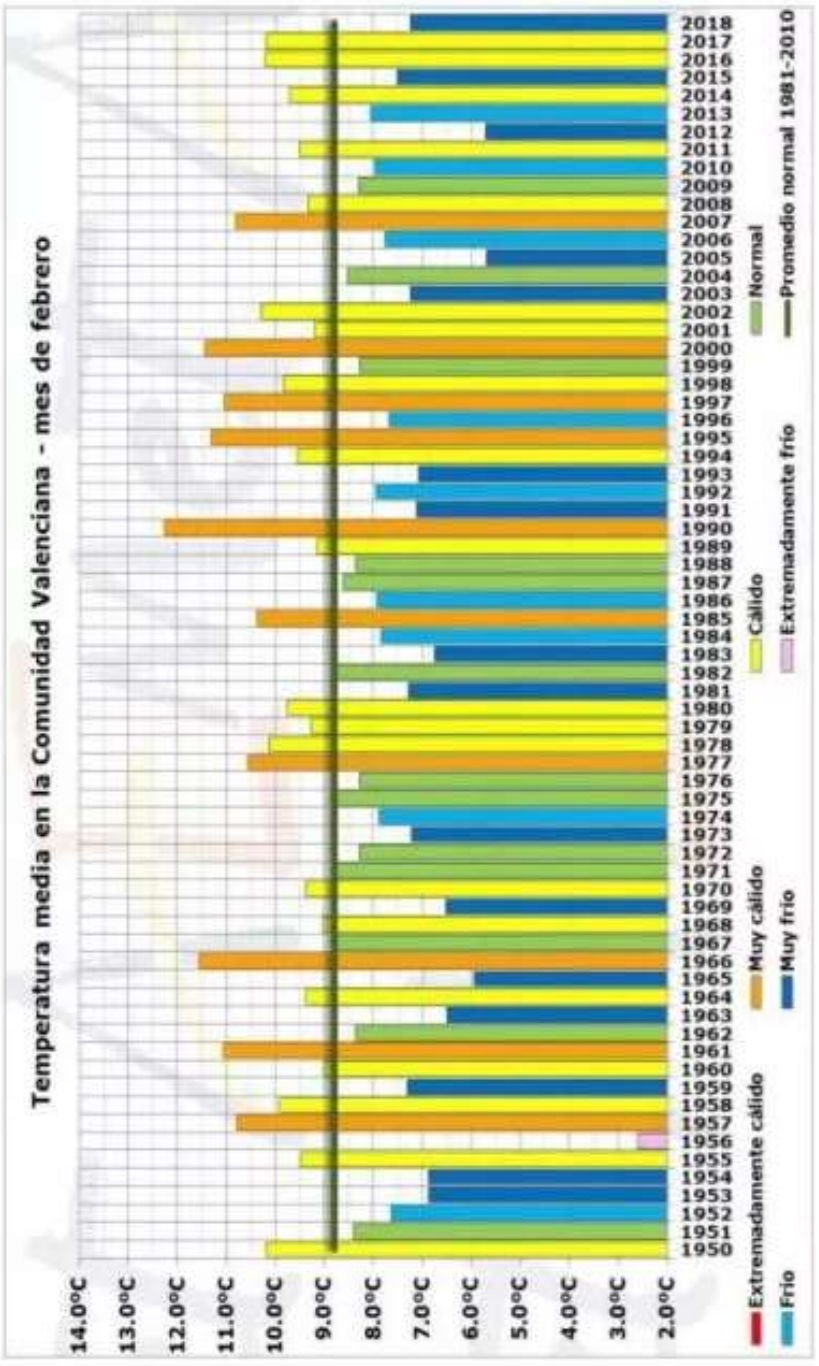

Figura 7.- Temperaturas media en la Comunidad Valencia de febrero 1950-2018 (Fuente: AEMET, 2019)

Debemos completar la tabla anterior con las temperaturas del mes de febrero del 2019, disponible en la Tabla 6, que al contener las temperaturas medias de varias estaciones meteorológicas habría que hallar la media de esos datos.

Tabla 6.- Temperaturas en valencia febrero 2019

Observatorio	Temperatura media (febrero de 2019)	Temperatura media Promedio normal (1981-2010)	Anomalía
Morella	8.3ºC	5.4ºC	+2.9ºC
Ontinyent	11.1ºC	9.7ºC	+1.4ºC
Utiel	7.7ºC	6.4ºC	+1.3ºC
Novelda	13.0ºC	11.8ºC	+1.2ºC
Vilafranca	6.1ºC	4.9ºC	+1.2ºC
Montserrat	12.8ºC	11.6ºC	+1.2ºC
Xàtiva	12.3ºC	11.1ºC	+1.2ºC
Segorbe	9.8ºC	8.9ºC	+0.9ºC
Atzeneta del Maestrat	9.6ºC	8.9ºC	+0.7ºC
Castelló	12.2ºC	11.7ºC	+0.5ºC
València	13.0ºC	12.5ºC	+0.5ºC
Elche/Ele	13.0ºC	12.7ºC	+0.3ºC
Benicarló	11.1ºC	10.9ºC	+0.2ºC
Sueca	13.1ºC	12.9ºC	+0.2ºC
Miramar	12.7ºC	12.8ºC	-0.1ºC
Rojales	12.3ºC	12.5ºC	-0.2ºC
Alicante/Alacant	11.9ºC	12.4ºC	-0.5ºC

MINISTERIO PARA LA TRANSICIÓN ECOLÓGICA

Agencia Estatal de Meteorología

Aplica el criterio del diagrama de Box-Whisker para comprobar si la temperatura media de febrero del 2019 es una temperatura atípica en la serie 1950-2019.

5 Algunas conclusiones

En la descripción estadística de un conjunto de datos, en ocasiones, aparecen ciertas complicaciones debido a la presencia de asimetría o a la existencia de atípicos. En esos casos, hay que poner en cuestión la información que proporciona la media aritmética, la cual puede dar una idea equivocada sobre la distribución de los datos. Sólo cuando exista bastante simetría y no aparezcan valores extremos influyentes, podemos dar como medida representativa de los datos la media aritmética y como medida de dispersión la desviación típica. En otros casos será más representativa como medida de tendencia central por ejemplo la mediana, y como medida de dispersión el intervalo y el rango intercuartílico.

Por otra parte, salvo que se tenga completa certeza de que un atípico detectado proviene de un error de medida o transcripción, la actitud frente a los atípicos será la de valorar el significado de su existencia e importancia en la interpretación global de los datos.

6 Bibliografía

AEMET-(2019). Avance climatológico de febrero de 2019 en la Comunidad Valenciana. Ministerio para la transición ecológica. Recuperado en: http://www.aemet.es/documentos/es/serviciosclimaticos/vigilancia_clima/resumenes_climat/ccaa/comunitat-valenciana/avance_climat_val_feb_2019.pdf

Brase, C.H.; Brase, C.P. (2016) *Understanding Basis Statistics. Metric Version*. Cengage Learning

Peck, R. (2014). *Statistics. Learning from data*. Books/Cole Cengage Learning.

Statista (2018-2019). Tope salarial de los equipos de fútbol de la 1ª división de la liga española para la temporada 2018/2019 (en millones de euros). Disponible en https://es.statista.com/estadisticas/498947/presupuesto-equipos-de-futbol-de-la-liga-en-espana/

Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K. (2012) *Probabilidad y estadística para ingeniería y ciencias*. Pearson.