M-estimators for interval-valued random elements

Beatriz Sinova Fernández

Universidad de Oviedo

Resumen del trabajo candidato para el Premio Ramiro Melendreras

XXXV Congreso Nacional de Estadística e Investigación Operativa

SEIO 2015

XXXV Congreso Nacional de Estadística e Investigación Operativa
IX Jornadas de Estadística Pública

S E
I O

Pamplona, 2015

Resumen del trabajo

Este trabajo se enmarca en el análisis estadístico con datos imprecisos, los cuales constituyen uno de los nuevos tipos de datos que van surgiendo al ampliarse aún más el campo de aplicaciones de la Estadística. El análisis de esta clase de datos presenta una serie de retos, debidos en buena parte a las peculiaridades del espacio en el que se trabaja. Este trabajo se ha centrado en el caso de los datos intervalares, cuyas aplicaciones al mundo real son muy numerosas en áreas tan diversas como la Economía, la Psicología, la Biomedicina, las Ingenierías, etc.

La teoría de los intervalos aleatorios (compactos), entendida como un caso especial de la teoría de conjuntos aleatorios compactos y convexos, ha impulsado la formalización matemática de muchos conceptos estadísticos sobre los mecanismos aleatorios que generan datos intervalares, así como el interés por el desarrollo de técnicas estadísticas que faciliten su análisis.

La filosofía perseguida en este trabajo es la de conservar, en la medida de lo posible, las nociones y técnicas clásicas utilizadas para variables y vectores aleatorios, si bien este enfoque no carece de obstáculos por las características propias de este tipo de datos. En este sentido, cabe reseñar la falta de linealidad del espacio de valores de esos datos (intervalos compactos no vacíos) con la aritmética usual y, como consecuencia de dicha semilinealidad, la inexistencia de una operación diferencia que simultáneamente esté siempre bien definida y conserve las propiedades que posee en el caso real en relación con la operación suma.

Como herramienta clave para el desarrollo o adaptación de técnicas estadísticas para intervalos aleatorios, se manejarán distancias entre intervalos compactos, evitando así los inconvenientes que con seguridad presentarían las diferencias entre intervalos.

Dentro de todo el abanico de hipotéticos estudios estadísticos con datos intervalares, el trabajo candidato al Premio Ramiro Melendreras se centra en la propuesta de medidas de localización robustas. Se trata de un objetivo de enorme interés, ya que en los últimos años se han ido ampliando exponencialmente los métodos estadísticos generalizados a intervalos aleatorios, pero la prática totalidad de ellos se basan en la media de Aumann. Si bien el valor de Aumann goza de abundantes propiedades, muy convenientes desde el punto de vista tanto estadístico como probabilístico, también hereda de la media de variables aleatorias reales su excesiva sensibilidad ante la existencia de datos atípicos (*outliers*) o los cambios en los datos.

Con el propósito de reforzar en este aspecto los métodos existentes en la literatura y cimentar adecuadamente los modelos aún por desarrollar, se vuelve acuciante la búsqueda de medidas que resuman la tendencia central de los intervalos aleatorios y que no se vean tan sumamente influenciadas por errores u observaciones atípicas como la media de Aumann. En pocas palabras, que salvaguarden las conclusiones estadísticas obtenidas a través de esos métodos, incluso bajo esos supuestos cambios con respecto a las condiciones iniciales o ideales para el estudio.

Estado del arte

En este apartado se resumirán las ideas que ya se habían recogido en la literatura acerca del problema planteado para este trabajo. Hay que tener en cuenta que muy pocas veces se había abordado la cuestión de estimar la localización de un intervalo aleatorio a través de un estimador que a su vez tomara valores de intervalo y, al mismo tiempo, fuera robusto. En este punto, sería

conveniente hacer una referencia a los dos tipos de vertientes que se originaron en el campo de los intervalos aleatorios.

- Por un lado, algunos expertos consideran que la obtención de datos intervalares está motivada por la imprecisión o los errores de medición a la hora de observar una variable aleatoria. En este caso, el objetivo está siempre ligado a la variable aleatoria subyacente. Entre dichos trabajos cabe citar, aunque para el caso más general de valores difusos, Ban *et al.* [1], Bodjanova [2], Grzegorzewski [3], Kersten [4] y Yamashiro [14, 15]. Nótese que la medida que examinan es la mediana y que no se complementan las propuestas con ningún estudio formal de la robustez.

- Por el otro lado, otros expertos centran su atención en el estudio de la los intervalos aleatorios que representan características esencialmente imprecisas. Si bien este último es el enfoque que impulsó esta investigación, conviene aclarar que los resultados obtenidos son aplicables independientemente de la concepción que se tenga sobre su significado e interpretación, siempre que las conclusiones estadísticas atañan a las características imprecisas y no a las posibles reales subyacentes.

Puede decirse, por lo tanto, que el estudio robusto completo de la localización de un intervalo aleatorio prácticamente se origina con la tesis doctoral de la autora de este trabajo, tanto en la amplia propuesta de medidas como en el uso de herramientas de la Teoría de la Robustez Estadística para fundamentar teóricamente el acierto de dichas alternativas y la mejora que suponen en comparación con la media de Aumann.

Aunque las medidas de localización aquí recogidas fueron introducidas en su momento en la tesis doctoral, presentada a finales de junio de 2014, es justo comentar que varios de los resultados más importantes y globalizadores aquí expuestos (como la medibilidad, la consistencia y el punto de ruptura de los M-estimadores de localización) se han demostrado con posterioridad a la defensa de la tesis. De todos modos, y como se va a detallar a continuación, existen varias publicaciones acerca de los conceptos *ad hoc* de tipo L^1 y alguna con un estudio parcial de la noción *ad hoc* de tipo L^2 (de ahí que no se hayan incluido demostraciones de los resultados ya publicados), pero el estudio general de los M-estimadores de localización para datos intervalares es completamente novedoso.

Como se ha ido referenciando a lo largo del trabajo, la propuesta de las extensiones de la noción de mediana al caso intervalar a través de métricas se encuentra ya en la literatura (véanse, por ejemplo, Sinova *et al.* [8, 10] y Sinova y Van Aelst [11, 13], y asimismo Sinova *et al.* [9, 12] para la situación más general de los números difusos aleatorios, del que los intervalos aleatorios son un caso particular).

Contribuciones de este trabajo

Con respecto a las nociones de mediana ya presentes en la literatura, este trabajo da un paso más allá al unificar todos esos conceptos *ad hoc* y otros novedosos desde el punto de vista de la literatura (los M-estimadores definidos a través del Teorema de Representación) bajo un mismo marco teórico: el de ***los M-estimadores de localización***.

El que los M-estimadores de localización de variables aleatorias reales sean considerados una herramienta exitosa en el estudio de la tendencia central, motiva sin duda su extensión al caso intervalar. Recuérdese que los M-estimadores se definen a través de un problema de minimización de cierta función de pérdida evaluada en las distancias euclídeas entre los valores que toma la variable aleatoria y la clase de números reales. Por lo tanto, la mediana es un caso particular de M-estimador de localización: el que se obtiene al escoger como función de pérdida la función valor absoluto. De hecho, la idea subyacente es la de proponer estimadores intermedios entre la media (con función de pérdida la función cuadrado) y la mediana (con función de pérdida la función valor absoluto).

Existen en la literatura algunos trabajos que adaptan los M-estimadores de localización al caso de elementos aleatorios que toman valores en un espacio de Hilbert, si bien sus autores las enmarcan únicamente en el contexto de la estimación robusta de la función núcleo de densidad (véanse Kim [5] y Kim y Scott [6, 7]). Gracias a la existencia de un encaje isométrico del espacio de intervalos compactos en el cono $\mathbb{R} \times [0, \infty)$ del espacio de Hilbert $\mathbb{R}^2$, es posible introducir los M-estimadores de localización para intervalos aleatorios.

Los teoremas de *existencia y unicidad de solución* (el Teorema de Representación), así como el estudio de la *solución algorítmica*, son aplicaciones de los logros publicados en los artículos anteriores por Kim y Scott. Una de las conclusiones más interesantes de la particularización del Teorema de Representación es que, bajo las condiciones del mismo (que cumplen muchas de las funciones de pérdida más notables en este marco), los M-estimadores pueden expresarse como combinaciones lineales convexas de los valores muestrales, lo que en el caso que se estudia da lugar a estimaciones con valores intervalares (gracias a que las las combinaciones convexas de sus imágenes en el cono en el que se encajan isométricamente permanecen en el mismo).

La contribución del trabajo presentado al Premio Ramiro Melendreras es el *análisis formal de la medibilidad, consistencia y robustez* de esta alternativa, así como el estudio completo del concepto *ad hoc* de tipo L^2. Las pruebas de dichos resultados son demasiado largas para la longitud permitida para el trabajo, así que en su mayor parte se han incluido simplemente indicaciones esquemáticas acerca de su demostración.

El estudio riguroso llevado a cabo en el trabajo se complementa con *estudios de simulación* también novedosos, en los que se comparan, desde el punto de vista de varios criterios estadísticos, algunos de los M-estimadores más notables obtenidos. Estos estudios de simulación conllevan una dificultad adicional, ya que no se dispone aún de modelos de distribuciones para intervalos aleatorios que sean suficientemente realistas como para ajustarse bien y representar un amplio espectro de situaciones prácticas.

Referencias

[1] Ban A, Coroianu L, Grzegorzewski P (2013) A fixed-shape fuzzy median of a fuzzy sample. *8th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2013, Advances in Intelligent Systems Research* **32**: 215–222

[2] Bodjanova S (2005) Median value and median interval of a fuzzy number. *Inform. Sci.* **172(1)**: 73–89

[3] Grzegorzewski P (1998) Statistical inference about the median from vague data. *Control and Cybernetics* **27**: 447–464

[4] Kersten PR (1995) The fuzzy median and the fuzzy MAD. In: *Proc. 3rd International Symposium on Uncertainty Modeling and Analysis and Annual Conference of the North American Fuzzy Information Processing Society, (ISUMA - NAFIPS'95)*, IEEE: pp. 85–88

[5] Kim JS (2011) *Kernel Methods for Classification with Irregularly Sampled and Contaminated Data.* PhD Thesis, University of Michigan (http://deepblue.lib.umich.edu/bitstream/handle/2027.42/89858/stannum_1.pdf?sequence=1)

[6] Kim JS, Scott CD (2011) On the robustness of kernel density M-estimators. In: *Proc. 28th Int. Conf. Mach. Lear.*, Bellevue, Washington: pp. 697–704

[7] Kim JS, Scott CD (2012) Robust kernel density estimation. *J. Mach. Learn. Res.* **13**: 2529–2565

[8] Sinova B, Casals MR, Colubi A, Gil MA (2010) The median of a random interval. In: *Combining Soft Computing and Statistical Methods in Data Analysis* (Borgelt C, González-Rodríguez G, Trutschnig W, Lubiano MA, Gil MA, Grzegorzewski P, Hryniewicz O, Eds), Adv. Intel. Soft Comp. Vol. 77. Springer, Berlin: pp. 575–583

[9] Sinova B, Gil MA, Colubi A, Van Aelst S (2012) The median of a random fuzzy number. The 1-norm distance approach. *Fuzzy Sets Syst.* **200**: 99–115

[10] Sinova B, González-Rodríguez G, Van Aelst S (2013) An alternative approach to the median of a random interval using an L^2 metric. In: *Sinergies of Soft Computing and Statistics for Intelligent Data Analysis* (Kruse R, Berthold MR, Moewes C, Gil MA, Grzegorzewski P, Hryniewicz O, Eds), Adv. Intel. Syst. Comp. Vol. 190. Springer, Berlin: pp. 273–281

[11] Sinova B, Van Aelst S (2013) Comparing the medians of a random interval defined by means of two different L^1 metrics. In: *Towards Advanced Data Analysis by Combining Soft Computing and Statistics* (Borgelt C, Gil MA, Sousa JMC, Verleysen M, Eds), Stud. Fuzz. Soft Comp. Vol. 285. Springer, Berlin: pp. 75–86

[12] Sinova B, Pérez-Fernández S, Montenegro M (2015) The wabl/ldev/rdev median of a random fuzzy number and statistical properties. In: *Strengthening Links Between Data Analysis and Soft Computing. Advances in Intelligent Systems and Computing* (Grzegorzewski P, Gagolewski M, Hryniewicz O, Gil MA, Eds), Adv. Int. Syst. Comp. Vol. 315. Springer, Heidelberg: pp. 143–150

[13] Sinova B, Van Aelst S (2015) On the consistency of a spatial-type interval-valued median for random intervals. *Statist. Probab. Lett.* **100**: 130–136

[14] Yamashiro M (1995) The median for a L-R fuzzy number. *Microelectronics Reliability* **35(2)**: 269–271

[15] Yamashiro M (1994) The median for a trapezoidal fuzzy number. *Microelectronics Reliability* **34(9)**: 1509–1511