Revista Colombiana de Estadística
Nº 12 - Diciembre - 1985

TERMINOLOGIA EN EL ESTUDIO DE VALORES EXTREMOS

Lina Sánchez T.

Profesora Asistente
Universidad Nacional

Resumen. Uno de los tópicos de la Estadística en donde más proliferación de términos hay para referirse a una sola expresión, es en el estudio de valores extremos. En este artículo se consideran algunas definiciones de *valor extremo*, *valor discordante*, *observación contaminante*, y *observación influyente*, con el fin de llegar a conclusiones sobre sus significados. También se considera el fenómeno de *enmascaramiento* y *empantanamiento* para establecer la diferencia entre su significado y sus efectos.

Abstract. One of the topics in Statistics where we see the greatest proliferation of terms referring to one single expresion, is the study of *outliers* or *spurious* observations. In this paper various definitions of the terms: *outlier, discordant, contaminant* and *influencial observations*, will be considered, in order to draw some conclusions about their meaning. The fenomena of *masking* and *swamping* will also be considered in order to establish the difference between their meaning and effects.

1. Introduccion.

Es cada vez más frecuente en los círculos culturales la realización de eventos que invitan a reflexionar sobre el lenguaje utilizado en diferentes áreas de conocimiento. Uno de ellos es el Congreso sobre la Terminología de la Edificación (realizado en Octubre de 1985) en el cual se abordaron entre otros temas como: Estudio de Lenguajes y Técnicas vernáculas mediante atlas linguístico, métodos de encuesta y análisis de textos: Terminología de las normas y las normas de la terminología; la normalización del Lenguaje Científico y Técnico; La definición y uso de términos técnicos. Si bien es cierto que estos temas son de especial interés para el mundo de la Arquitectura (por en-

focarse el congreso en esta disciplina), ellos abren una perspectiva para realizar intentos similares en el área de la Estadística.

Uno de los tópicos de la Estadística en donde existen más proliferación de términos para referirse a un solo concepto es en el estudio de *Valores Extremos* presentes en una muestra.

La falta de estandarización en la terminología, es un serio inconveniente para lograr una efectiva comunicación en los ensayos escritos y evitar así confusión en algunos conceptos. Puede ser que este problema se ha acentuado al haber desarrollado una gran cantidad de literatura sobre el tema (una muestra de ellos son los 200 trabajos publicados de 1978 a 1983) basada en fuentes bibliográficas que cubren dos siglos de estudio, sin hacer intentos de unificar los conceptos y la terminología.

Solo hace tres años se publicó un estudio donde se critican algunas expresiones usadas en el tratado de valores extremos (Barnett 1983).

Pero la confusión existente puede ser más bien síntoma de un problema más serio: la falta de claridad y acuerdo sobre algunos conceptos. Este es el punto de vista explorado en el presente artículo, de allí que se hace referencia a

expresiones claves en el tratado de valores extremos, se presenta diferentes acercamientos a su significado y se obtienen conclusiones al respecto.

2. Significado de términos usados en el estudio de Valores Extremos.

Se considera en primer lugar el desarrollo del significado de la palabra *outlier*, luego se establecen las diferencias entre *outlier*, *observación discordante*, *contaminante* e *influyente*.

Ya en 1777 Bernoulli comenta sobre la práctica de descartar lo que él llama valores discordantes, práctica que data de 200 años atrás! En 1838, Bessel resalta en un trabajo de geodecia que él nunca rechaza una observación solamente por el tamaño "grande" de su residuo; se debe permitir que todas las observaciones contribuyan al resultado final.

De allí en adelante surgen criterios para el rechazo adecuado de valores que uno no esperaría encontrar en una muestra. Por ejemplo, después de presentar una discusión acerca de los problemas que surgen cuando ocurren "valores extremos" dentro de lecturas astronómicas, Wright (1884)

sugiere que la mejor regla es rechazar cualquier observación cuyo residual excede en magnitud 5 veces al error probable, es decir, 3.37 veces la desviación estándar. La razón que expone para ello, es que si se cumple la ley Gaussiana, únicamente más o menos una observación en 1000 será rechazada y "se hará muy poco daño en cualquier caso".

Estos acercamientos al tema, muestran que la visión de dichos autores se basa en situaciones donde la variabilidad intrínseca y los errores ordinarios de medida o de ejecución resultan de un diseño donde el patrón de variación es muy cercano al normal, donde se teme que de vez en cuando una observación sea afectada por un error grande.

Entonces, se ha definido *outlier* como una observación con un residual "anormalmente grande". Outlier se traduce como *valor extremo*. Otros términos usados para referirse a una observación de tal naturaleza son: valor resagado, no formal, super-valor, valor discordante, anormal, aberrante, contrahecho, falso, ilegítimo, discrepante, contaminante, influyente, dudoso o sospechoso. Estas expresiones señalan el hecho de que la observación así llamada se encuen

tra localizada muy lejos del resto de las observaciones en estudio, de tal forma que a los ojos del investigador llega a ser dudosa o sorprendente.

Considérence las siguientes definiciones con el fín de establecer diferencias entre algunos de los términos más comunes para referirse al término *outlier*:

Observación Discordante. Es toda observación que parece sorprendente o discrepante al investigador (Beckman and Cook (1983)).

Observación Contaminante. Toda observación que no proviene de la población que se está estudiando, sino de otra población. (Jain (1981)).

Observación Influyente. Una observación es influyente, si al no tenerla en cuenta para hacer el análisis de datos, se alteran sustancialmente rasgos importantes de dicho análisis (específicamente en el análisis mínimo cuadrático de los datos basado en un modelo de regresión lineal) (Cook (1979)).

Valor Extremo. Una observación con un residual anormalmente grante. (Anscombe (1960)).

La confusión aparece cuando algunos autores de ensayos sobre éste tópico, usan indistinta-

mente las expresiones anteriores. Para mencionar solo algunos casos, se presentan las siguientes situaciones:

i) En un artículo en el que se analiza un trabajo de Beckman y Cook (1983) sobre Valores Extremos, se hace la siguiente afirmación: "... Sarnett y Lewis (1978, p.22) usan el término "outlier" cuando los autores usan "observaciones discordantes". Pero Barnett y Lewis (p.23) usan "observaciones discordantes" cuando nuestros autores usan "Contaminantes "!". (McCulloch-Meeter (1983)).

ii) En un estudio sobre "outliers" (Cook (1983)), se presenta la siguiente definición; Outlier: una expresión para referirse a ambas una observación contaminante o una observación discordante.

La misma traducción de outlier es ya una expresión propensa a crear confusión, pues un *valor extremo*, no necesariamente es un *outlier*; o sea en una muestra ordenada de valores $x_1 < x_2 < \ldots < x_n$, x_1 y x_n son los valores extremos de la muestra, pero ello no implica que sus residuales sean "anormalmente grandes".

Al usar el concepto de *valor discordante*,

se está suponiendo de antemano la existencia de un modelo de referencia, por ejemplo el normal (fig.1), en donde x_n es un valor extremo y sorprendente con respecto a los valores de la variable que representa el modelo.

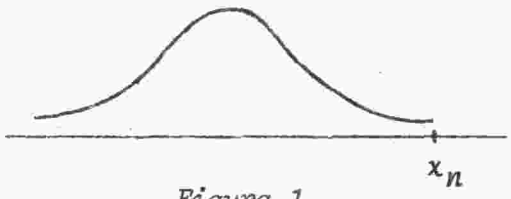

Figura 1

x_n es un valor *discordante* para este caso, podría decirse que es un valor extremo *discordante*.

La noción de valor *contaminante* tiene que ver con el mecanismo que origina los valores con residuales muy grandes. Estos valores pueden generarse cuando los datos provienen de dos poblaciones o distribuciones; una de ellas llamada la distribución básica, genera los valores "buenos"; mientras que la otra, llamada distribución contaminante, genera los valores contaminantes. Esta situación se puede representar por medio de los siguientes modelos de contaminación:

en una muestra proveniente de una población normal A, con media μ y varianza σ^2

i) una o más observaciones provienen de una distribución B, $N(\mu+\lambda, \sigma^2)$ $\lambda > 1$ (figura 2).

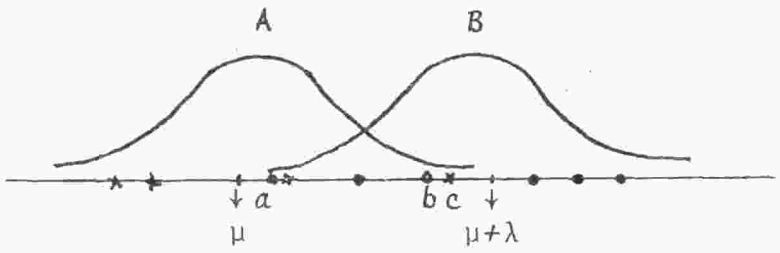

Figura 2

lo cual es un tipo de contaminación que genera un error de localización.

ii) una o más observaciones provienen de una distribución C, $N(\mu, \lambda^2\sigma^2)$, $\lambda > 1$ (figura 3). Este tipo de contaminación genera errores escalares.

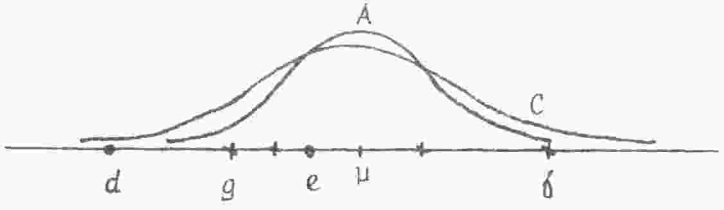

Figura 3

+ observaciones de A
o observaciones de B
o observaciones de C

Examinando las posibles situaciones plantea das en los gráficos, se puede concluir que aunque una observación extrema puede ser *contaminante*, por ejemplo las observaciones *b* y *d*, un valor *contaminante no necesariamente es un valor extremo*, por ejemplo las observaciones *a* y *e*. A su vez, un valor *extremo puede ser no contaminante*, como son las observaciones c y *f*.

La observación *a*, es una observación proveniente de otra distribución que no es la distribución objeto de estudio, *a*, contamina a la muestra obtenida de la población A, pero muy probablemente no luciría "sorprendente" a los ojos del investigador; lo mismo puede afirmarse de la observación *e* de la figura 3. Un valor *contaminante puede no ser discordante*.

Si una observación es extrema en el sentido de tener un residual anormalmente grande, dicha observación es *discordante*.

En cuanto a las observaciones influyentes consideradas como aquellas que ejercen una marcada influencia en los estimadores mínimo cuadráticos del vector de parámetros en regresión lineal, puede afirmarse que una observación *influyente es una que es extrema*, bajo el punto de vista de su influencia sobre el análisis que se

lleva a cabo en la muestra.

3. Consideraciones respecto a los fenómenos de Enmascaramiento y Empantanamiento.

Se ha definido el fenómeno de enmascaramiento (masking) como la pérdida de potencia de las pruebas aplicadas sucesivamente para detectar valores extremos múltiples (esto es, en muestras donde hay mínimo dos valores extremos).

Sin embargo, esta definición apunta más al efecto que se produce cuando el fenómeno de enmascaramiento está presente que a lo que realmente es dicho fenómeno.

El *enmascaramiento* se presenta cuando una muestra contiene varios valores extremos, los cuales incrementan de tal forma la expansión de la muestra, que el quitar uno de estos valores, produce un pequeño mejoramiento en la apariencia de la muestra; particularmente todos los valores $|x_i-\bar{x}|/s$ estan cerca a cero dado el valor grande de la desviación estándar s.

En conclusión el fenómeno de *enmascaramiento* ocurre cuando un valor extremo no es declarado como tal porque otro valor extremo, el más cercano a él, encubre su importancia.

Al presentar esta situación, la aplicación de pruebas sucesivas para detectar uno a uno tales valores, es infructuosa y se requiere de la aplicación de pruebas más sofisticadas para descubrirlos. La consecuencia es que la prueba utilizada para tal fin pierde potencia al ser aplicada sucesivamente y tiende a declarar menos valores extremos de los que realmente hay.

El fenómeno de *empantamiento* (swamping), se ha definido como aquel que se produce cuando la prueba para detectar valores extremos múltiples en aplicaciones sucesivas, tiende a declarar más observaciones extremas de las que realmente hay.

También esta definición hace referencia a los efectos que se presentan cuando el fenómeno está presente. El *empantamiento* ocurre cuando al aplicar una prueba sucesivamente para detectar valores extremos múltiples, un valor extremo altamente discordante arrastra consigo a otro valor que es inofensivo. De allí que la prueba declare más valores extremos de los que realmente hay en la muestra. Por ejemplo, puede ser que el valor extremo d (Fig.3) arrastre consigo el valor g, el cual no es extremo y la prueba aplicada declare dos valores como extremos cuando en realidad hay uno.

Si se consideran las definiciones y planteamientos presentados anteriormente, se concluye que los términos *valor extremo, discordante, contaminante* e *influyente* describen conceptos diferentes y por lo tanto no pueden ser usados indistintamente.

Es imporante y necesario adoptar una terminología estándar, con el fín de evitar la proliferación de términos y facilitar tanto la realización de futuros trabajos, como el entendimiento del tema por parte del lector. También será de gran beneficio si cada autor incluye en su estudio las definiciones de las expresiones que aparecerán en él.

* *

BIBLIOGRAFIA

Anscombe, F.J., (1960). Rejetion of Outliers. Technometrics, Vol. 2 # 2.

Barnett, V., (1978). Outliers in Statistical Data. John Wiley, New York.

Barnett, V., (1933). Discussion. Technometrics. Vol. 25 # 2.

Beckman, R.J. and Cook, R.D., (1983). Outliers. Technometrics, Vol. 25 # 2.

Cook, R.D., (1979). Influencial Observations in Linear Regression. Journal of the American Statistical Association, Vol. 74 # 365.

Hawkins, D.M., (1980). Identifications of Outliers. Printing House University, Cambrige.

Jain, R.B., (1981). Detecting Outliers. Commun. Statist. - Theor. Meth. Lancaster.

McCulloch, C. and Meeter, D. (1983). Discussion, Techsometrics, Vol. 25 # 2.

* *