Revista Colombiana de Estadística
Nº 28

# DETECCION DE UN OUTLIER SUPERIOR EN MUESTRAS EXPONENCIALES BASADO EN PREDICCION DE LA MAYOR OBSERVACION

DAIRO S. GIL G.
Escuela de Matemáticas y Estadística Universidad
Pedagógica y Tecnológica de Colombia

JOSÉ ALBERTO VARGAS N.
Departamento de Matemáticas y Estadística Universidad Nacional de Colombia.

RESUMEN. En este trabajo se presentan dos estadísticas para detectar un outlier superior en muestras exponenciales, cuando se utiliza el estimador de Kaminsky para estimar la mayor observación. Por medio de simulación se comparan los resultados con los obtenidos por Balasooriya y con los resultados obtenidos sin el uso del predictor.

## INTRODUCCION

Últimamente se ha escrito mucho acerca de los valores *sorpresivos*, aquellos que se apartan del grueso de las observaciones en una muestra. Este fenómeno ha llamado la atención de los investigadores a través de los tiempos. En comentarios hechos por Bernoulli en 1777 se señala que la práctica de descartar las observaciones discordantes era común desde hacía 200 años. El primer intento por desarrollar métodos estadísticamente objetivos para el problema de outliers fue reportado alrededor de 1850 según Beckman y Cook (1983). Sin embargo, el concepto de outlier aparece vago hoy día como lo fue hace 200 años. Es común que los investigadores los nombren de manera diferente. Así, comúnmente, aparecen reportados en la literatura, como observaciones discordantes, valores extremos, observaciones anómalas, contaminantes, valores sorpresivos, valores sucios, outliers, para mencionar sólo algunos términos que se han citado a través de los años. Collett y Lewis (1976) reportan los resultados de un experimento para investigar la naturaleza subjetiva de la decisión para catalogar una observación como outlier, y concluyen que la disposición para catalogar una observación como outlier depende del método de presentación (aleatorio, ordenado o

gráfico), de la experiencia y de la escala de los datos. Las observaciones extremas tienden a parecer más discrepantes en la medida en que la escala sea más grande; además, cuando los tamaños de muestra son moderados, es fácil realizar una inspección visual de los datos para detectar outliers; pero en conjuntos más grandes o más complicados como en el caso de la Regresión, muestras multivariadas, Diseño de Experimentos, etc., la inspección visual provechosa de los datos puede resultar completamente imposible. Luego, resulta necesario aplicar cualquier tipo de criterio que sea objetivo, al conjunto de datos, para juzgar la presencia de outliers sin descartar la inspección visual de ellos.

Cuando se enfrenta el problema de los outliers, debe siempre tenerse presente la posibilidad de que se dé el efecto de enmascaramiento que es la tendencia que tienen observaciones extremas, a esconder o enmascarar el efecto de observaciones más extremas que si son realmente outliers Vargas (1992). Sin importar el énfasis que se ha hecho sobre la manera subjetiva como los elementos de una muestra pueden ser declarados outliers, debe tenerse en mente que hay familias de distribuciones que son susceptibles de producirlos.

En el presente trabajo se proponen dos estadísticas para detectar un outlier superior cuando la muestra proviene de una distribución Exponencial. Se calculará el porcentaje de veces que cada estadística detecta un outlier superior en muetras exponenciales, cuando se utiliza el predictor de kamynsky y, finalmente, se compararán los resultados con los reportados por Balasooriya (1989) y con los obtenidos sin el uso del predictor.

## Estadística para detectar un outlier superior en muestras exponenciales.

Kaminsky y Nelson (1975), dan el mejor predictor lineal insesgado para el $s$-ésimo estadístico de orden $\vec{X}_{(s)}$ basado en $X_{(1)}, X_{(2)}, \ldots, X_{(r)}$

$$\vec{X}_{(s)} = x_{(r)} + \delta(r,s)\vec{\theta}$$

donde

$$\delta(r,s) = \sum_{j=r+1}^{s} (n-j+1)^{-1}$$

y

$$\vec{\theta} = \frac{\sum_{i=1}^{s} x_{(j)} + (n-r)x_{(r)}}{r}$$

es el estimador de máxima verosimilitud de $\theta$ . En la clase de todos los estimadores que son funciones de los primeros r estadísticos de orden, $\vec{\theta}$ es además el estimador insesgado de $\theta$ de varianza mínima. Balasooriya (1989) utiliza el estimador de kaminsky para proponer la estadística

$$W_r = \frac{x_{(r+1)} - x_{(r)}}{\vec{x}_{(r+1)} - x_{(r)}}$$

que permite probar la hipótesis, $H_0$: Todas las $n$ observaciones provienen de una distribución exponencial con parámetro de escala $\theta$ desconocido contra la alternativa $H_1$: las observaciones, en el conjunto de datos, provienen de distribuciones exponenciales con parámetros de escala $\theta_i = C_i\theta$, $i = 1, 2, \ldots, n$, donde los $C_i$ son constantes positivas y al menos $m$ de ellas son diferentes de la unidad. Usualmente $m$ es un número pequeño y su valor exacto puede ser desconocido. La función de densidad de probabilidad $g_1$ de $W_r$ es:

$$g_1(w_r) = \frac{r^{r+1}}{(r + w_r)^{r+1}}, \quad w_r > 0$$

Y la correspondiente función de distribución acumulativa es:

$$G_1(w_r) = 1 - \frac{r^r}{(r + w_r)^r}, \quad w_r > 0$$

Balasooriya (1989) simula las potencias de las pruebas para $n = 10$, $n = 15$ y $n = 20$ Un miembro fijo de cada muestra fue contaminado multiplicándolo por un predeterminado valor c para crear un solo outlier en el conjunto de datos.

En este trabajo se proponen las siguientes dos estadísticas:

$$T_{1a} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)}}$$

y

$$T_{2a} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)} - x_{(1)}}$$

para probar la presencia de un sólo outlier superior cuando la muestra proviene de una distribución exponencial, de parámetros a y $\theta$ , y para probar la presencia de un solo outlier superior, cuando la muestra proviene de una distribución exponencial con parámetro $\theta$ independiente del origen, respectivamente.

## Cálculo de valores críticos y potencias.

Se considerarán las siguientes cuatro estadísticas:

$$T_1 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)}}$$

$$T_2 = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$$

$$T_{1a} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)}}$$

$$T_{2a} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)} - x_{(1)}}$$

Para cada una de las estadísticas $T_{1a}$ y $T_{2a}$ se procedió a calcular los valores críticos al 1% y al 5% de la siguiente manera:

i) Se generaron 1000 muestras aleatorias de tamaño n de una distribución Exponencial con parámetro de escala $\theta = 1$ y para cada muestra se estima la n-ésima observación $\vec{x}_{(n)}$, basada en las $n-1$ observaciones anteriores, utilizando el estimador de Kaminsky

$$\vec{x}_{(n)} = \frac{n x_{(n-1)}}{n-1} + \frac{\sum_{j=1}^{n-1} x_{(j)}}{n-1}$$

ii) Para cada muestra se calculan los valores

$$T_{1ac} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)}}$$

$$T_{2ac} = \frac{x_{(n)} - x_{(n-1)}}{\vec{x}_{(n)} - x_{(n-1)}}$$

iii) Se ordenan por separado los valores de $T_{1ac}$ y $T_{2ac}$, de menor a mayor, y se encuentran los percentiles 95 y 99, que serán los valores críticos simulado al 5% y al 1%, respectivamente, para cada estadística.

iv) Debido a que para el cálculo de cada estadística y para cada valor crítico se utilizan muestras diferentes, es natural que se observe cierta variabilidad en los valores críticos cuando se repite el proceso. Por tal razón, y para lograr estabilidad en los valores críticos obtenidos, cada procedimiento se iteró 20 veces y se tomó como valor crítico el promedio de los valores críticos de las 20 iteraciones.

v) Los pasos anteriores se realizaron para muestras de $n = 10$, $n = 15$ y $n = 20$. Para el algoritmo computacional se utilizó el lenguaje Turbo Pascal y el comando Randomize del mismo.

La tabla 1 muestra los valores críticos para cada estadística y para cada tamaño de muestra. En ella se incluyen, además, los valores críticos al 1% y al 5% para las pruebas con las estadísticas $T_1$ y $T_2$ obtenidos de las tablas Barnet and Lewis (1984) p 771 y para la estadística $W_r$.

Para cada una de las estadísticas $T_1$, $T_2$, $T_{1a}$ y $T_{2a}$ se calcula el porcentaje de veces que la estadística detecta un outlier superior generado mediante la contaminación de

**TABLA 1 : Valores críticos al 1% y al 5% para cada tamaño de muestra y para cada una de las estadísticas**

| ESTADÍSTICA | NIVEL $\alpha$ % | $n_1 = 10$ | $n_2 = 15$ | $n_3 = 20$ |
|---|---|---|---|---|
| $T_{1,\alpha}$ | 1 | 0.7680 | 0.7150 | 0.6820 |
| * | 5 | 0.6580 | 0.6010 | 0.5670 |
| $T_{2,\alpha}$ | 1 | 0.7830 | 0.7240 | 0.6870 |
| * | 5 | 0.6750 | 0.6100 | 0.5730 |
| $T_{1a,\alpha}$ | 1 | 2.0634 | 1.6795 | 1.4758 |
| ** | 5 | 1.2582 | 1.0438 | 0.9144 |
| $T_{2,\alpha}$ | 1 | 2.1276 | 1.7277 | 1.4818 |
| ** | 5 | 1.2982 | 1.0654 | 0.9185 |
| $W_{r,\alpha}$ | 1 | 6.0130 | 5.4529 | 5.2112 |
| *** | 5 | 3.5550 | 3.3404 | 3.2450 |

* Valores obtenidos de las tablas de valores críticos Barnet y Lewis (1984) p 771.

** Valores calculados con base en 1000 muestras generadas por el método de Monte Carlo.

*** Valores obtenidos a partir de la distribución acumulativa de $W_r$, Balasooriya (1989).

la observación mayor, para tres tamaños de muestra y dos niveles de significancia, de la siguiente manera:

i) Se generan 1000 muestras aleatorias de tamaño n de una distribución Exponencial con parámetro de escala $\theta = 1$ y para cada muestra ordenada se estima la $n$-ésima observación usando el estimador de Kaminsky

ii) Para cada muestra se calculan los valores

$$T_{1C} = \frac{Cx_{(n)} - x_{(n-1)}}{Cx_{(n)}}$$

$$T_{2C} = \frac{Cx_{(n)} - x_{(n-1)}}{Cx_{(n)} - x_{(1)}}$$

$$T_{1aC} = \frac{Cx_{(n)} - x_{(n-1)}}{\vec{x}_{(n)}}$$

$$T_{2aC} = \frac{Cx_{(n)} - x_{(n-1)}}{\vec{x}_{(n)} - x_{(1)}}$$

Para $C = 1, 2, 3, 5$ donde $C$ es el contaminante ($C = 1$ significa que la muestra está libre de contaminación).

iii) Se ordenan separadamente los valores calculados de cada $T_{1C}$, $T_{2C}$, $T_{1ac}$ y $T_{2ac}$ de menor a mayor y para cada tamaño de muestra $n$, cada nivel de contaminación y cada nivel de significancia $\alpha$.

iv) Se calcula el porcentaje de detección para cada estadística, a cada valor de contaminación, para cada tamaño de muestra y para cada nivel de significancia $\alpha$.

v) Los pasos anteriores se realizan para muestras de $n = 10$, $n = 15$ y $n = 20$, para cada nivel de contaminación $C = 1$, $C = 2$, $C = 3$, $C = 5$ y para cada nivel de significancia 1% y 5%. Para el algoritmo computacional se utilizó el lenguaje Turbo Pascal y el comando Randomize del mismo.

La tabla 2 contiene los porcentajes de detección obtenidos. Se incluyen los resultados reportados por Uditha Balasooriya (1989) p 716 .

TABLA 2 : Porcentaje calculado de pruebas significativas basadas en 1000 muestras generadas por el método de Monte Carlo.

| ESTADÍSTICA | $\alpha$ | $n$=10 | | | | $n$=15 | | | | $n$=20 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | c=1 | c=2 | c=3 | c=5 | c=1 | c=2 | c=3 | c=5 | c=1 | c=2 | c=3 | c=5 |
| $T_1$ | 1 | 1.4 | 15.1 | 45.8 | 100.0 | 0.7 | 25.1 | 69.3 | 4100.0 | 90.5 | 26.1 | 87.3 | 100.0 |
| | 5 | 5.1 | 44.1 | 100.0 | 100.0 | 4.9 | 59.4 | 100.0 | 100.0 | 93.9 | 68.7 | 100.0 | 100.0 |
| $T_2$ | 1 | 0.9 | 15.7 | 43.2 | 100.0 | 1.0 | 18.8 | 67.1 | 4100.0 | 91.1 | 23.4 | 89.7 | 100.0 |
| | 5 | 5.3 | 42.6 | 98.8 | 100.0 | 5.0 | 56.7 | 100.0 | 100.0 | 94.3 | 68.0 | 100.0 | 100.0 |
| $T_{1a}$ | 1 | 1.2 | 17.7 | 51.3 | 100.0 | 1.4 | 22.0 | 71.5 | 4100.0 | 91.2 | 26.0 | 93.9 | 100.0 |
| | 5 | 5.3 | 43.6 | 99.7 | 100.0 | 5.1 | 57.5 | 100.0 | 100.0 | 94.7 | 68.7 | 100.0 | 100.0 |
| $T_{2a}$ | 1 | 1.2 | 18.9 | 51.8 | 100.0 | 0.8 | 25.0 | 72.5 | 4100.0 | 91.4 | 29.0 | 94.7 | 100.0 |
| | 5 | 6.5 | 48.4 | 100.0 | 100.0 | 5.1 | 58.0 | 100.0 | 100.0 | 96.6 | 70.5 | 100.0 | 100.0 |
| $W_r$ | 1 | 1.1 | 18.5 | 53.2 | — | 0.7 | 25.2 | 76.1 | — | 90.4 | 31.5 | 88.3 | — |
| | 5 | 4.5 | 48.0 | 96.1 | — | 3.9 | 64.8 | 99.9 | — | 94.4 | 72.8 | 100.0 | — |

—Indica que Balasooriya no presentó resultados para ese nivel de contaminación.

Indica que Balasooriya no presentó resultados para ese nivel de contaminación.

Después de observar los resultados, se puede comentar que, por ejemplo, la estadística $T_{1a}$ detecta el outlier un 99.7% de las veces a un nivel de significancia del 5%, cuando se multiplica la mayor observación de la muestra por 3, en tanto que la de Balasooriya lo hace en el 96.1%. A ese mismo nivel de significancia, la estadística $T_{2a}$ lo detecta el 100% de las veces.

En general, las estadísticas prouestas son potentes para detectar un outlier superior en una muestra exponencial. Aún para n=10 con c=3, las dos estadísticas han detectado el outlier en 99.7% de las muestras o más. Claramente, son más potentes

con valores grandes de $n$. La estadística $T_{2a}$ es, en general, levemente más potente que $T_{1a}$.

## 4. EJEMPLOS

Ejemplo 1 : La siguiente muestra ordenada de tamaño 10 se generó a partir de una Exponencial con $\theta = 1$ en un programa en Turbo Pascal usando el comando Randomize: 0.08960922, 0.2723378, 0.5192777, 0.5256515, 0.6092543, 0.6167258, 1.047988, 1.194195, 1.200829, 3.722681.

La décima observación ordenada fué entonces multiplicada por 2 para crear un outlier en el conjunto de datos. Para probar la hipótesis nula $H_0$: la muestra no contiene un outlier superior, contra la alternativa de un lado. La tabla 3 contiene los valores calculados para cada una de las estadísticas y los respectivos valores críticos.

| | $T_1$ | $T_2$ | $T_{1a}$ | $T_{2a}$ | $w_9$ |
|---|---|---|---|---|---|
| Valor Calculado | 0.8387 * | 0.8489 * | 3.1077 * | 3.2528 * | 7.7234 * |
| $\alpha = 1\%$ | 0.7680 | 0.7830 | 2.0634 | 2.1276 | 6.0130 |
| $\alpha = 5\%$ | 0.6580 | 0.6750 | 1.2582 | 1.2982 | 3.5550 |
| $\widehat{x}_{(10)} \longrightarrow$ | | | | | 2.009350924 |

* Valor significativo al 1 y al 5 porciento. Observe que la prueba resulta significativa al 1% y al 5% en cada caso

Ejemplo 2 : Los siguientes son los datos reportados por Proschan sobre la duración de equipos de aire acondicionado instalados en aviones para ilustrar una aplicación general de la prueba. La exponencialidad de los datos de la muestra fué previamente establecida por Stephens (1978) usando, no sólo, el procedimiento de prueba de Shapiro and Wilk (1972) sino su versión modificada: 14, 27, 32, 34, 54, 57, 59, 61, 66, 67, 102, 134, 152, 209, La tabla 4 muestra los resultados de las pruebas.

Tabla 4 :Valores calculados y valores críticos de las estadísticas para probar un outlier superior.

| | $T_1$ | $T_2$ | $T_{1a}$ | $T_{2a}$ | $w_9$ |
|---|---|---|---|---|---|
| Valor Calculado * | 0.0913 | 0.0972 | 0.06995 | 0.073375 | 0.23023 |
| $\alpha = 1\%$ | 0.7680 | 0.7830 | 2.0634 | 2.1276 | 6.0130 |
| $\alpha = 5\%$ | 0.6580 | 0.6750 | 1.2582 | 1.2982 | 3.5550 |
| $\widehat{x}_{(10)} \longrightarrow$ | | | | | 300.2142857 |

Valor Calculado * Ninguna prueba resultó significativa.

## 5. CONCLUSIONES

Después de observar el comportamiento de las estadísticas propuestas se puede comentar que son, por lo menos, igualmente potentes que la estadística de Balasooriya y en algunos casos más potentes.

Los resultados observados indican que el hecho de utilizar el estimador de Kaminsky para estimar el valor máximo, mejora la capacidad de detección del outlier superior con respecto a $T_1$ y $T_2$, lo que es más evidente cuando se usa el nivel de significancia del 5%.

Para muestra de tamaño grande aumenta en forma significativa la potencia, como era de esperarse .

Al igual que las otras estadísticas consideradas, éstas pueden ser afectadas por el efecto de enmascaramiento, lo cual valdría la pena de ser investigado.

Por todo lo que se ha podido ver en el transcurso de este trabajo, es claro que las estadísticas propuestas permiten detectar un outlir superior cuando la distribución es Exponencial y que el uso del estimador de kamisky mejora la capacidad de detectarlo.

## BIBLIOGRAFÍA


Balasooriya, U (1989), *Detection of outliers in the exponential distribution based on prediction*, Commun. Statist.- Theory Meth. **18(2)**, 711-719.

Barnett, V. and Lewis, T. (1984), *Outliers in statistical data*, Segunda edicción, John Wiley, New York.

Beckman, R. J. and Cook, R. D. (1983), *Outlier...s(with Discussion)*, Technometrics **25**, 119-149.

Collett, D. and Lewis. T. (1976), *The subjective nature of outlier rejection procedures*, Applied Statistics **25**, 228-237.

Kaminsky, K. S. and Nelson, P. I. (1975), *Best linear umbiased prediction of order statistics in location and scale*, Amer. Statist. Assoc. **70**, 145 - 150.

Neyman, J. and Scott, E. L. (1971), *Outlier proness of phenomena and of related distribution*, In Rustagi, J. (Ed) (1971)., *Optimising method in statistics*, vol. 43, Academic Press, New York.

Shapiro, S. S. and Wilk, M. B. (1972), *An Analysis of Variance Test for the Exponencial Distribution*, Technometric **14**, 355-370.

Stephens, M. A. (1978), *On the W-test for exponenciality whit origen Known*, Technometric **20**, 33-35.

Vargas, J. A. (1993), *Outliers en muestras exponenciales censuradas*, Revista Colombiana de Estadística **27**, 1-9.