Revista Colombiana de Estadística

No.35

# DETECCIÓN DE DOS O MÁS VALORES ATÍPICOS SUPERIORES EN MUESTRAS ALEATORIAS EXPONENCIALES UTILIZANDO LA TÉCNICA DE PREDICCIÓN DE LAS OBSERVACIONES MÁXIMAS

CARLOS PANZA OSPINO[a] Y JOSÉ ALBERTO VARGAS[b]

[a] Profesor, Universidad del Atlántico.
[b] Profesor Asociado, Universidad Nacional de Colombia.

RESUMEN Partiendo de investigaciones realizadas por *Balasooryia* (1989) y Gil y Vargas (1993), en el siguiente trabajo se prueba que tres estadísticas existentes aumentan ligeramente su potencia para detectar dos o tres valores atípicos superiores en muestras aleatorias exponenciales, si son modificadas con el predictor lineal de Kaminsky.
PALABRAS CLAVES: Distribución exponencial, Predictor lineal o estimador de Kaminsky, Valores atípicos, Test de discordancia.

## 1. INTRODUCCIÓN

Las observaciones que parecen apartarse del contexto de una serie de datos se denominan observaciones discordantes o valores atípicos. Los primeros intentos para desarrollar métodos estadísticamente objetivos en el tratamiento y manejo de tales observaciones datan de mediados del siglo pasado. En la actualidad existen dos métodos para el tratamiento de valores atípicos. El primero de ellos, basado en la utilización de los tests de discordancia, propugna por la identificación de los valores atípicos con el fin de descartarlos del todo o de incorporarlos al estudio para obtener información importante del fenómeno bajo estudio. El segundo de los métodos hace referencia a la posibilidad de acomodar un valor atípico mediante una modificación

apropiada del modelo y/o del método de análisis. En cierta medida, por razones obvias, los métodos de identificación son preferidos a los de acomodación. Todo lo señalado hasta ahora , proporciona las bases necesarias para proponer un procedimiento de detección de más de un valor atípico tomando como punto de partida muestras aleatorias exponenciales, toda vez que en el medio que nos circunda son innumerables los fenómenos relacionados con este tipo de distribución.

## 2.DEFINICIÓN DEL PROBLEMA

En general, la cantidad de valores atípicos en un conjunto de datos no se conoce de antemano y evidentemente éstos pueden aparecer en cualquier parte de una muestra ordenada. Sin embargo, en muchas aplicaciones se espera que sólo acurran un número reducido de tales valores en la muestra. Sean $X_{(1)}, X_{(2)}, ..., X_{(n)}$ las observaciones de una muestra aleatoria proveniente de una distribución exponencial con una función de densidad

$$f(x, \lambda) = \lambda \exp(-\lambda x); \quad \lambda > 0, \; x > 0$$

Para una muestra dada de $n$ observaciones se quiere contrastar la hipótesis nula $H_0$ : las $n$ observaciones provienen de una misma distribución exponencial, contra la hipótesis alternativa $H_1$ : las observaciones provienen de una distribución exponencial con parámetro de escala $\lambda_i = C_i \lambda$ donde $i = 1, 2, ..., n$ y al menos $m < n$ constantes positivas $C_i$ son diferentes de la unidad.

El presente trabajo propone dos estadísticas para la detección de un par de valores atípicos superiores y una estadística para la detección de $m > 2$ valores atípicos superiores, basadas todas ellas en la predicción de las observaciones máximas de

una muestra cuando las primeras $r \geq 1$ observaciones en ella no representan valores atípicos . Se utiliza en las estadísticas propuestas el predictor lineal de Kaminsky como elemento modificador de tres estadísticas para tests de discordancia estudiadas en Barnet y Lewis (1984).

Se proponen la estadística

$$\frac{x_{(n)} - x_{(n-2)}}{\hat{x}_{(n)}}$$

para detectar un par de valores atípicos superiores $x_{(n-1)}$ y $x_{(n)}$ en una muestra proveniente de una distribución exponencial con origen conocido; la estadística

$$\frac{x_{(n)} - x_{(n-2)}}{\hat{x}_{(n)} - x_{(1)}}$$

para detectar un par de valores atípicos superiores $x_{(n-1)}$ y $x_{(n)}$ en una muestra proveniente de una distribución exponencial con origen desconocido; la estadística

$$\frac{x_{(n)} + \ldots + x_{(n-k+1)}}{\sum_{1}^{n-k} x_{(j)} + \hat{x}_{(n-k+1)} + \ldots + \hat{x}_{(n)}}$$

para detectar $k\,(\geq 2)$ valores atípicos superiores $x_{(n-k+1)}, \ldots, x_{(n)}$ en una muestra proveniente de una distribución exponencial con origen conocido. En el presente estudio se comparan los resultados del poder detector de las estadísticas mencionadas con los obtenidos mediante las estadísticas que no utilizan el modificador.

En todas las estadísticas propuestas $\hat{x}_{(*)}$ es la estimación de la estadística de orden correspondiente mediante el predictor lineal de Kaminsky. Para una muestra de tamaño $r$ ordenada ascendentemente, Kaminsky (1975) y Kaminsky y Nelson

(1977) dedujeron el mejor predictor lineal insesgado del $s$-ésimo estadístico de orden cuando se conocen los $r$ primeros. La expresión matemática del predictor lineal tiene la forma

$$\hat{x}_{(s)} = x_{(r)} + \delta(r,s)\hat{\theta}$$

donde

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

y $\hat{\theta}$ es el estimador máximo verosímil, con varianza mínima, del parámetro $\theta$ de la distribución de la cual provienen los $r$ estadísticos de orden y viene dado por

$$\hat{\theta} = \frac{\sum_{1}^{r} x_{(j)} + (n-r)\,x_{(r)}}{r}$$

## 3. OBTENCIÓN DE VALORES CRÍTICOS

Los valores críticos de las estadísticas que se proponen en este trabajo se obtendrán por simulación utilizando para ello el procedimiento de Monte Carlo. Como no se tenían valores críticos de las estadísticas homólogas que no involucran el modificador, hubo también necesidad de simularlos. El proceso de obtención de los valores críticos de las estadísticas se describe a continuación.

Se generan 1000 muestras aleatorias ordenadas, de igual tamaño, provenientes de una distribución exponencial con párametro de escala $\lambda = 1$ y para cada una de ellas se estiman las dos observaciones superiores como se indica abajo utilizando para el

efecto el predictor lineal de Kaminsky:

$$\hat{x}_{(n-1)} = \frac{(n-1)\,x_{(n-2)} + \sum_{1}^{n-2} x_{(i)}}{n-2}$$

$$\hat{x}_{(n)} = \frac{(n+1)\,\hat{x}_{(n-1)} + \sum_{1}^{n-2} x_{(i)}}{n-1} \qquad (3.1)$$

Una vez hechas las estimaciones anteriores, para cada muestra se calculan las siguientes estadísticas:

$$T_1 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)}}$$

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

$$T_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}$$

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

$$T_3 = \frac{x_{(n)} + x_{(n-1)}}{\sum_{1}^{n} x_{(j)}}$$

$$T_{3m} = \frac{x_{(n)} + x_{(n-1)}}{\sum_{1}^{n-2} x_{(j)} + \hat{x}_{(n-1)} + \hat{x}_{(n)}}$$

Para las estadísticas que detectan tres valores atípicos se estiman las tres observa-

ciones superiores de cada muestra ordenada de la siguiente manera:

$$\hat{x}_{(n-2)} = \frac{(n-2)\,x_{(n-3)} + \sum_{1}^{n-3} x_{(i)}}{n-3}$$

$$\hat{x}_{(n-1)} = \frac{n\hat{x}_{(n-2)} + \sum_{1}^{n-3} x_{(i)}}{n-2} \qquad (3.2)$$

$$\hat{x}_{(n)} = \frac{(n+1)\,\hat{x}_{(n-1)} + \sum_{1}^{n-3} x_{(i)} + \hat{x}_{(n-2)}}{n-2}$$

y se calculan las estadísticas

$$T_4 = \frac{x_{(n)} + x_{(n-1)} + x_{(n-2)}}{\sum_{1}^{n} x_{(i)}}$$

$$T_{4m} = \frac{x_{(n)} + x_{(n-1)} + x_{(n-2)}}{\sum_{1}^{n-3} x_{(i)} + \hat{x}_{(n-2)} + \hat{x}_{(n-1)} + \hat{x}_{(n)}}$$

A continuación se ordenan ascendentemente y por separado los valores calculados de cada una de las estadísticas anteriores. Para cada estadística se hallan los percentiles 0.95 y 0.99, los cuales determinan los valores críticos del 5% y 1%, respectivamente. Como es natural, se observó cierta variabilidad en la obtención de valores críticos de una misma estadística para cada conjunto diferente de 1000 muestras aleatorias. Por tal razón, el procedimiento descrito se iteró 20 veces y en calidad de valor crítico se tomó el promedio de los valores arrojados por cada iteración. Cabe aquí anotar

que se calcularon los valores críticos de una misma estadística en conjuntos de 2000 muestras aleatorias con 20 iteraciones del proceso y de 4000 muestras aleatorias con una sola iteración y los resultados obtenidos no se diferenciaron significativamente de los obtenidos con 1000 muestras.

Los pasos descritos se realizaron en muestras de tamaño 10, 15 y 20, para lo cual fue necesaria la elaboración de un algoritmo computacional. La tabla 1 contiene los valores críticos simulados para cada estadística.

## 4. ESTUDIO DE POTENCIA

Para cada una de las estadísticas $T_*$ y $T_{*m}$ se calcula el porcentaje de veces en que cada una de ellas detecta la presencia de dos o tres valores atípicos superiores generados previamente por la contaminación de las observaciones máximas. El procedimiento se realiza para tres tamaños de muestra y dos niveles de significación tal y como se explica a continuación.

Se generan 1000 muestras aleatorias de igual tamaño, provenientes de una distribución exponencial con párametro de escala $\lambda = 1$ y para cada muestra ordenada se estiman las dos (tres) observaciones superiores utilizando el predictor lineal de Kaminsky como se indica en las fórmulas (3.1) y (3.2). Luego se contamina cada muestra multiplicando los valores de sus dos (tres) observaciones superiores por una misma constante $C$ que toma cualquiera de los valores 1,2,3 o 5. La constante $C$ se conoce con el nombre de contaminante. Si $C = 1$ se dice que la muestra está libre de contaminación.

**TABLA 1 Valores críticos de las estadísticas estudiadas para tamaños de muestra y niveles de significancia estipulados**

| ESTADÍSTICA | NIVELES DE SIGNIFICANCIA $\alpha$, % | TAMAÑOS DE MUESTRA | | |
|---|---|---|---|---|
| | | n=10 | n=15 | n=20 |
| $T_1$ | 5 | 0.768 | 0.704 | 0.665 |
| | 1 | 0.849 | 0.788 | 0.749 |
| $T_{1m}$ | 5 | 1.405 | 1.134 | 1.008 |
| | 1 | 2.360 | 1.770 | 1.485 |
| $T_2$ | 5 | 0.786 | 0.714 | 0.670 |
| | 1 | 0.861 | 0.797 | 0.756 |
| $T_{2m}$ | 5 | 1.452 | 1.155 | 1.024 |
| | 1 | 2.429 | 1.800 | 1.501 |
| $T_3$ | 5 | 0.643 | 0.505 | 0.418 |
| | 1 | 0.716 | 0.571 | 0.476 |
| $T_{3m}$ | 5 | 0.848 | 0.586 | 0.463 |
| | 1 | 1.183 | 0.770 | 0.575 |
| $T_4$ | 5 | 0.771 | 0.623 | 0.526 |
| | 1 | 0.821 | 0.681 | 0.579 |
| $T_{4m}$ | 5 | 1.008 | 0.697 | 0.559 |
| | 1 | 1.375 | 0.892 | 0.689 |

(*)Los valores críticos de la estadística $T_3$ están tabulados en Barnett y Lewis (1984) p.337.Los valores críticos de las demás estadísticas fueron simulados.

Para cada muestra contaminada se calculan los valores de las estadísticas detectoras, tanto de las modificadas como de las no modificadas, comparando cada uno de los resultados obtenidos para una misma estadística con los respectivos valores críticos simulados al 1% y 5%. Si el valor observado de la estadística es mayor que el valor crítico, entonces se puede afirmar que la estadística en mención ha detectado la presencia de los dos (tres) valores atípicos generados en la muestra. Seguidamente se cuenta el número de casos en los cuales una misma estadística detecta los valores atípicos superiores generados por contaminación. La potencia de la estadística viene

dada por la frecuencia relativa de detección de valores atípicos superiores.

Los pasos anteriores de realizaron en muestras de tamaño 10 , 15 y 20 , en cada una de las cuales se tiene en cuenta cuatro niveles de contaminación , a saber: $C = 1, C = 2, C = 3$ y $C = 5$. El algoritmo computacional fue realizado en el programa Microsoft EXCEL. La tabla 2 contiene los porcentajes de detección obtenidos.

## 5. EJEMPLOS

**5.1 Ejemplo uno.** Las ocho primeras observaciones de la siguiente muestra aleatoria de tamaño $n = 10$ fueron simuladas y provienen de una distribución exponencial con párametro de escala $\lambda = 1$. Las dos últimas observaciones son valores atípicos especialmente generados en la muestra por contaminación.

La muestra contaminada es

| | | | | |
|---|---|---|---|---|
| 0.00517097 | 0.05620920 | 0.16651045 | 0.44033846 | 0.53092432 |
| 0.56079144 | 0.59516881 | 0.60697032 | 1.93127100 | 5.92618255 |

Todas las estadísticas calculadas para esta muestra arrojaron resultados significativos al 1% y 5%; por lo tanto, se puede afirmar que las estadísticas detectaron la presencia de los valores atípicos contaminantes.

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

| ESTADISTICA | NIVELES DE SIGNIFICANCIA $\alpha$, % | TAMAÑO DE MUESTRA | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n=10 | | | | n=15 | | | | n=20 | | | |
| | | NIVELES DE CONTAMINACIÓN | | | | | | | | | | | |
| | | C=1 | C=2 | C=3 | C=5 | C=1 | C=2 | C=3 | C=5 | C=1 | C=2 | C=3 | C=5 |
| $T_1$ | 5 | 4,8 | 39,3 | 77,9 | 100,0 | 6,0 | 50,4 | 95,8 | 100,0 | 3,4 | 57,5 | 100,0 | 100,0 |
| | 1 | 0,6 | 12,4 | 37,0 | 86,2 | 0,8 | 14,9 | 58,6 | 100,0 | 0,4 | 26,5 | 73,9 | 100,0 |
| $T_{1m}$ | 5 | 5,0 | 39,5 | 78,6 | 100,0 | 5,2 | 52,7 | 95,9 | 100,0 | 4,2 | 60,1 | 99,8 | 100,0 |
| | 1 | 0,5 | 13,0 | 38,5 | 87,7 | 0,6 | 19,7 | 62,5 | 100,0 | 0,8 | 27,9 | 80,6 | 100,0 |
| $T_2$ | 5 | 4,7 | 36,6 | 75,2 | 100,0 | 5,5 | 49,9 | 95,5 | 100,0 | 3,6 | 56,8 | 100,0 | 100,0 |
| | 1 | 0,9 | 11,7 | 35,0 | 82,9 | 0,8 | 19,0 | 57,0 | 100,0 | 0,3 | 25,9 | 72,6 | 100,0 |
| $T_{2m}$ | 5 | 4,9 | 39,3 | 78,6 | 100,0 | 5,2 | 53,1 | 95,8 | 100,0 | 4,3 | 59,8 | 99,8 | 100,0 |
| | 1 | 0,6 | 13,4 | 38,5 | 87,4 | 0,6 | 19,7 | 62,3 | 100,0 | 0,9 | 28,5 | 81,0 | 100,0 |
| $T_3$ | 5 | 4,9 | 55,3 | 90,2 | 100,0 | 6,2 | 69,2 | 98,3 | 100,0 | 3,5 | 76,4 | 99,4 | 100,0 |
| | 1 | 0,4 | 20,5 | 62,0 | 98,4 | 1,2 | 35,2 | 83,4 | 100,0 | 0,5 | 42,4 | 93,3 | 100,0 |
| $T_{3m}$ | 5 | 4,5 | 54,3 | 94,9 | 100,0 | 5,9 | 73,0 | 98,7 | 100,0 | 4,3 | 81,3 | 99,9 | 100,0 |
| | 1 | 0,4 | 20,4 | 63,1 | 99,9 | 0,5 | 34,7 | 88,0 | 100,0 | 0,7 | 48,3 | 97,4 | 100,0 |
| $T_4$ | 5 | 4,9 | 52,9 | 88,1 | 100,0 | 5,4 | 72,6 | 98,7 | 100,0 | 4,4 | 80,5 | 99,8 | 100,0 |
| | 1 | 1,0 | 21,8 | 64,3 | 97,8 | 1,4 | 39,5 | 86,8 | 100,0 | 0,6 | 50,3 | 96,4 | 100,0 |
| $T_{4m}$ | 5 | 4,4 | 54,0 | 93,1 | 100,0 | 5,4 | 75,9 | 99,9 | 100,0 | 4,6 | 87,9 | 100,0 | 100,0 |
| | 1 | 0,9 | 21,0 | 65,1 | 99,7 | 0,7 | 40,2 | 92,9 | 100,0 | 0,7 | 56,9 | 98,4 | 100,0 |

**5.2 Ejemplo dos.** Las observaciones ordenadas de la siguiente muestra de tamaño $n = 20$ representan una transformación logarítmica de los tiempos de espera, medidos en minutos, debido a las interrupciones de energía en el circuito de San Salvador de Barranquilla. La estimación de los párametros de confiabilidad para el proceso de Weibull exige que las observaciones transformadas provengan de una distribución exponencial. La exponencialidad de la muestra fue establecida mediante las pruebas de Kolmogorov-Smirnov y Shapiro-Wilks descritas en Peña (1988). los datos fueron facilitados por Jorge Sagre, ingeniero eléctrico de CORELCA, y hacen parte de su trabajo de grado para optar al título de Especialista en Estadística. La muestra es la siguiente

| | | | | |
|---|---|---|---|---|
| 0.034764991 | 0.086465703 | 0.124235558 | 0.125683629 | 0.135966288 |
| 0.166717660 | 0.167636794 | 0.170081669 | 0.267291143 | 0.329721827 |
| 0.485524695 | 0.528421005 | 0.529616032 | 0.573432712 | 0.574400104 |
| 0.853942832 | 0.853942832 | 0.858014886 | 0.865443923 | 2.380679589 |

Ninguna de las estadísticas calculadas para la muestra resultó significativa; por lo tanto, no existe suficiente evidencia para suponer que existan valores atípicos en dicha muestra.

## 6.CONCLUSIONES

Del trabajo realizado se puede concluir que las estadísticas modificadas son ligeramente más potentes que las respectivas homólogas sin modificador, aun en las muestras más pequeñas. Cuando el nivel de contaminación es alto y el tamaño de

muestra aumenta, todas las estadísticas estudiadas son igualmente potentes al 5% de nivel de significancia. Bajo esas mismas condiciones, las estadísticas modificadas con el estimador de Kaminsky son más potentes al 1% de nivel de significancia. Cuando no existen contaminantes en la muestras, las estadísticas modificadas reflejan un mejor comportamiento frente a los niveles de significancia, en cuanto a la detección errónea de valores atípicos, de allí que pueda resultar más conveniente su utilización especialmente cuando el tamaño de muestra no es muy grande. Como se ve, es posible aumentar la certidumbre en la detección de dos (tres) valores atípicos si para ello se utilizan las estadísticas modificadas con el estimador de Kaminsky.

## REFERENCIAS

Balasooryia, U.(1989). "Detection of outliers in the exponencial distribution based on prediction ". Commun. Statist. -Theory Meth. 18(2); 711-719.

Barnett, R and Lewis, T. (1984). "Outliers in statistical data. " 2a. ed. John Wiley, New York. 2-43; 144-161

Gil, D y Vargas J.A. (1993). "Detección de un outlier superior en muestras exponenciales basada en la predicción de la mayor observación" Revista Colombiana de Estadística 28; 37-42.

Kaminsky, K.S. (1975). "Bests linear prediction of order statistics in exponencial and Pareto populations " ARL Report, No.75-0201, Applied Mathematics Research laboratory. United States Air Force.

Kaminsky, K.S. and Nelson, P.I. (1977) "Best linear unbiased prediction of order statistics in location and scale families. " J. Amer. Statist. Assoc. 70; 145-150.

Peña Sánchez de Rivera, D. (1988) "Estadística. Modelos y Métodos. " Alianza editorial, Madrid, España.