Revista Colombiana de Estadística
Nº 27

OUTLIERS EN MUESTRAS EXPONENCIALES CENSURADAS

JOSÉ ALBERTO VARGAS N.

Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia

ABSTRACT. Some considerations about the presence of outliers in censored exponential samples are shown in this paper. A test statistic for detecting an upper outlier in a Type I censored exponential sample is proposed and its power when the outlier occurs anywhere in the sample is studied. It is also shown that this test suffers from the masking effect.

KEYWORDS. Failure times, Truncation time, Masking effect.

1. INTRODUCCION

1.1 Manejo de Outliers.

El problema de outliers ha recibido mucha atención durante las últimas decadas. Dos métodos para el manejo de outliers se mencionan en la literatura: Acomodación e Identificación.

El primer método utiliza procedimientos estadísticos para obtener inferencias válidas sobre la población los cuales no son muy afectados por la presencia de outliers. Estos procedimientos se dice que son robustos a la presencia de outliers y usualmente otorgan menos importancia a los valores extremos que a los otros miembros de la muestra. Por ejemplo, los estimadores L y M son a menudo usados con distribuciones simétricas.

El segundo método consiste en identificar los outliers para estudiarlos posteriormente; Beckman y Cook (1983) señalan que la identificación de un outlier puede conducir a: (a) su eliminación, (b) nueva información contenida en variables concomitantes que de otra forma pasaría inadvertida, (c) su incorporación a través de un nuevo modelo, o (d) revisión de los métodos experimentales. De cualquier manera, la opinión de una persona conocedora del problema debe ser tenida muy en cuenta para rechazar o retener un outlier.

Usualmente, el problema de identificación de outliers es considerado como uno de prueba de hipótesis. La hipótesis nula es que la muestra proviene de una distribución F, mientras que la hipótesis alternativa es que una o varias de las observaciones proviene de una distribución G. F y G pueden ser dos distribuciones completamente especificadas, o pueden ser dos miembros de una familia de distribuciones. Bajo la hipótesis nula, se calculan valores críticos de una estadística, y luego se comparan con los valores de la estadística en los problemas bajo investigación.

1.2 Efecto de Enmascaramiento.

El efecto de enmascaramiento es la tendencia que tienen observaciones extremas no declaradas como outliers, a esconder o enmascarar el efecto de observaciones más extremas que si son realmente outliers. Bendre y Kale (1985) introdujeron una medida del efecto de enmascaramiento como la pérdida de potencia debido a la presencia de un número mayor al de outliers bajo prueba. El método es el siguiente: Sean $X_1, \ldots, X_n$ variables aleatorias independientes. Bajo la hipótesis nula, estas variables aleatorias son idénticamente distribuídas, provenientes de una población con función de distribución F, mientras que bajo la hipótesis alternativa, los outliers aparecen de una población con función de distribución G. Las funciones de distribución F y G tienen la misma forma pero el parámetro de localización o escala de G, está modificado por una cantidad desconocida a. La hipótesis alternativa se puede escribir de la siguiente forma H_1: Una de las $X_i's\,(i = 1, \ldots, n)$ tiene función de distribución G. Sea T(X) una estadística utilizada para detectar un outlier, con región crítica $A_{n,\alpha}$. El comportamiento de la prueba se estudia con respecto a las siguientes medidas:

$$P_1(a) = P[T(X) \in A_{n,\alpha} \mid H_1]$$

y

$$P_2(a) = P[T(X) \in A_{n,\alpha} \mid \text{ más de una observación tiene distribución G}]$$

$P_1(a)$ es la potencia de la prueba, y $P_2(a)$ es la potencia de la prueba cuando más de un outlier está presente. Una medida del efecto de enmascaramiento se define como:

$$M_a = P_1(a) - P_2(a),$$

y el efecto de enmascaramiento límite como $M = \lim_{a \to a_0} M_a$, donde a_0 es el valor discordante límite. Una prueba sufre de enmascaramiento si M es positivo, y se dice que es libre de enmascaramiento si M es cero.

1.3 Muestras Exponenciales Censuradas.

La distribución exponencial es muy útil en situaciones donde se requiere conocer el tiempo de vida de una unidad o un sistema. En pruebas de vida, un número fijo de objetos, digamos n, son puestos a prueba simultáneamente. Los tiempos en los que los objetos fallan, y que se denominarán tiempos de falla, $x_1, x_2, \ldots$ son entonces registrados utilizando un plan de entre los siguientes posibles planes:

i) Muestreo completo: El experimento se continúa hasta que todos los n tiempos de falla hayan sido observados.

ii) Censuramiento Tipo I: El experimento termina en un tiempo fijo T.

iii) Censuramiento Tipo II: El experimento termina cuando el r-ésimo tiempo de falla x_r haya sido observado.

iv) Censuramiento mixto: Combinación de los dos anteriores. Es decir, el experimento termina cuando se complete el tiempo T, o el r-ésimo tiempo de falla x_r se observe.

La distribución exponencial ha sido frecuentemente utilizada como modelo de tiempos de vida, (ver por ejemplo, Bartholomew (1963), Lawless (1982) y Zacks (1986) entre otros). Ademas, debido a razones económicas o de tiempo, las muestras deben usualmente ser censuradas. Si a esto se añade la presencia de posibles outliers, el análisis de estas muestras presenta serios problemas.

Aunque la presencia de outliers en muestras censuradas es un problema que se presenta regularmente, solo un reciente artículo, Kimber (1990), presenta una metodología para su identificación. Kimber define las cantidades $I_S = 2(C_S - M)$ e $I_I = 2(M - C_I)$, dónde C_S , C_I son los cuartiles superior e inferior respectivamente y M es la mediana. Con datos censurados, los cuartiles son estimados mediante el uso del estimador Kaplan-Meier. El método propuesto por Kimber consiste en identificar como posible outlier, cualquier observación que se localice fuera del intervalo $(C_I - 1.5I_I, C_S + 1.5I_S)$. Este método, que puede ser usado con distribuciones asimétricas, tambien permite detectar outliers inferiores.

En este artículo se estudia el problema de un outlier superior en muestras exponenciales que deben ser terminadas en cierto tiempo T, fijado con anterioridad, y que se denomina tiempo de truncamiento.

2. LA ESTADISTICA W

Sea X_i una variable aleatoria con función de densidad

$$f(x;\theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta}) \ , x \geq 0, \theta > 0$$

Sea n el número de objetos puestos a prueba, T el tiempo en el cual termina el experimento y N (el cual se asume mayor que cero) el número de fallas observado en el intervalo $[0, T]$. N es una variable aleatoria binomial con parámetros n y $1 - \exp(-T/\theta)$.

Sean $Y_1, Y_2, \ldots, Y_n$ las variables aleatorias definidas por :

$$Y_i = \begin{cases} X_i & \text{si } X_i \leq T \\ T & \text{si } X_i > T \end{cases}$$

para $i = 1, \ldots, n$, y $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ las correspondientes estadísticas de orden. Para detectar un outlier superior, se propone la estadística W, definida de la siguiente manera:

$$W = \frac{Y_{(n)}}{\sum_{i=1}^{n} Y_{(i)}}.$$

Un valor grande de W provee evidencia para rechazar H_0, donde H_0 establece que la muestra no contiene outliers. Sin pérdida de generalidad se asume que $\theta = 1$. $P(W > a)$ se puede expresar de la siguiente manera:

(2.1)

$$P(W > a) = P(W > a \mid N = n)P(N. = n) + P(W > a \mid 0 < N < n)P(0 < N < n)$$

para $\frac{1}{n} \leq a \leq 1$.

El primer término en (2.1) considera el caso $N = n$. Esto significa que no hay observaciones censuradas. Por tanto, la distribución de $Y_i, (i = 1, \ldots, n)$ dado que $Y_i \leq T$ tiene la forma exponencial truncada. El segundo término en (2.1) calcula probabilidades cuando $0 < N < n$. Esta situación ocurre si hay por lo menos una observación censurada. En este caso, W se puede escribir como:

$$W = \frac{T}{\sum_{i=1}^{N} Y_i + (n - N)T}.$$

Vargas e Iglewicz (1992) demostraron que bajo H_0, (2.1) está definida de la siguiente forma:

$$\begin{aligned} P(W > a) = & \left[1 - \exp(-T/a) \sum_{i=0}^{n-1} \frac{(T/a)^i}{i!}\right] \\ & \times \sum_{j=1}^{[1/a]} \binom{n}{j} (-1)^{j-1} (1 - ja)^{n-1} - \sum_{j=1}^{[1/a]} \binom{n}{j} (-1)^{j-1} \exp(-jT) \\ & \times \left[1 - \exp(-\frac{T}{a}(1 - ja)) \sum_{i=0}^{n-1} \frac{(\frac{T}{a}(1 - ja))^i}{i!}\right] + \sum_{N=1}^{n-1} \binom{n}{N} \sum_{i=0}^{N} \binom{N}{i} \\ & \times (-1)^i \exp(-T(n - N + i)) \int_0^u p(\chi^2_{2N})\, d\chi^2_{2N} \end{aligned} \tag{2.2}$$

dónde $[1/a]$ es la parte entera de $1/a$, $u = 2\max\left(0, \frac{T}{a} - T(n - N + i)\right)$ y el integrando es la función de densidad de una variable aleatoria chi-cuadrado con 2N grados de libertad.

Valores críticos de W, calculados directamente de (2.2) para ciertos valores de T y n se presentan en la Tabla 1. Esta tabla presenta dos columnas adicionales con los valores críticos de W cuando las muestras no son censuradas. Estos últimos valores fueron reproducidos de Barnett y Lewis (1984) (Table I, p. 369-370).

Tabla 1 Valores críticos de W

Tamaño de Muestra	Tiempo de truncamiento	Nivel de significancia		Muestras no Censuradas	
n	T	5%	1%	5%	1%
10	8.0	0.4445	0.5338	0.4450	0.5358
	9.0	0.4448	0.5351		
	10.0	0.4449	0.5356		
15	8.0	0.3341	0.4037	0.3346	0.4069
	9.0	0.3345	0.4058		
	10.0	0.3346	0.4065		
20	8.0	0.2698	0.3257	0.2705	0.3297
	9.0	0.2703	0.3283		
	10.0	0.2704	0.3293		
30	9.0	0.1977	0.2395	0.1980	0.2412
40	9.0	0.1573	0.1895	0.1576	0.1915
50	9.0	0.1312	0.1573		
60	9.0	0.1129	0.1326	0.1131	0.1371
70	9.0	0.0993	0.1158		
80	9.0	0.0888	0.1029		
90	9.0	0.0805	0.0926		
100	9.0	0.0736	0.0842		
120	9.0	0.0630	0.0713	0.0632	0.0759

Comparando las entradas de cada columna en la Tabla 1, se observa que para valores grandes de T, los valores críticos para muestras censuradas son muy cercanos a los valores críticos en muestras no censuradas. Esto es debido a que cuando el valor de T aumenta, los resultados se aproximan al caso $T \to \infty$, que es exactamente la situación de muestras no censuradas.

3. ESTUDIO DE POTENCIA

Como se dijo anteriormente, en pruebas de identificación de outliers se examinan dos hipótesis. La hipótesis nula establece en general, que la muestra $X_1, \ldots, X_n$ proviene de una distribución F, mientras que la hipótesis alternativa establece que al menos una observación proviene de una distribución G. La potencia de la prueba bajo estudio, depende por consiguiente de la hipótesis alternativa que se considere. Varios modelos generadores de outliers han sido propuestos en la literatura para representar la situación anterior. Barnett y Lewis (1984) asi como Gather y Kale (1981) los han examinado. Los modelos mas conocidos son:

i) Modelo de outliers identificados: Aquí se asume que el conjunto de índices de las variables provenientes de G está identificado. Exactamente, la hipótesis alternativa H_I es: $X_{i_1}, X_{i_2}, \ldots, X_{i_{n-k}}$ tienen distribución F y $X_{i_{n-k+1}}, \ldots, X_{i_n}$ tienen distribución G, donde el conjunto $I_k = \{i_{n-k+1}, \ldots, i_n\}$ se asume que es conocido.

ii) Modelo de k outliers: Aquí la hipótesis alternativa H_k es similar a H_I excepto que el conjunto de índices I_k no es conocido, sino que puede ser cualquiera de los $\binom{n}{k}$ posibles conjuntos.

iii) Modelo de outliers extremos: Introducido por Barnett y Lewis (1984), este modelo enfatiza el papel que desempeñan los extremos de un conjunto de datos como los únicos posibles outliers. En este caso, la hipótesis alternativa para probar k outliers superiores se formula de la siguiente manera: $H_L : X_{(1)}, X_{(2)}, \ldots, X_{(n-k)}$ son estadísticas de orden provenientes de la distribución F, y $X_{(n-k+1)}, \ldots, X_{(n)}$ son estadísticas de orden de G. Aunque los outliers pueden aparecer en cualquier lugar de la muestra y no necesariamente en los extremos, este modelo toma en consideración el hecho, de que usualmente se sospecha la presencia de outliers luego de examinar los extremos.

Aunque la selección del modelo ha generado cierta discusión, los tres modelos mencionados han sido utilizados en diversas oportunidades. Si se ha definido una estadística para identificar un outlier superior por ejemplo, el primer modelo identifica la observación proveniente de G, el segundo modelo establece que la observación proveniente de G puede ser cualquiera de las n observaciones que componen la muestra, mientras que el tercer modelo establece que $x_{(n)}$ es la obsevación seleccionada de G.

El siguiente estudio de simulación se llevó a cabo: 5000 muestras de tamaño $n = 10$, 15 y 20 fueron seleccionadas de una distribución exponencial con parámetro 1.0. Una observación fué seleccionada aleatoriamente de cada muestra y multiplicada por un

número c, para varios valores de c, obteniendo $n-1$ observaciones de una población exponencial con parámetro 1.0 y una única observación de una población exponencial con parámetro c. La muestra fué truncada en cierto tiempo fijo T, y el valor de W fué calculado para cada muestra.

Toda la generación de muestras y los respectivos análisis fueron hechos usando un programa FORTRAN el cual incorporó subrutinas del IMSL (International Mathematical and Statistical Library). La Tabla 2 muestra las proporciones de rechazos de la hipótesis nula H_0 por W, con un nivel de significancia $\alpha = 0.05$. Se observa que la prueba es potente para valores grandes de T.

Tabla 2 Potencias de W con $\alpha = 0.05$

Tamaño de Muestra	Tiempo de Truncamiento	c			
n	T	2.0	5.0	10.0	$\to \infty$
10	7.0	0.085	0.211	0.261	0.524
	8.0	0.091	0.268	0.327	0.677
	9.0	0.098	0.305	0.380	0.794
	10.0	0.099	0.330	0.450	0.879
15	7.0	0.085	0.209	0.273	0.546
	8.0	0.095	0.280	0.378	0.727
	9.0	0.099	0.334	0.421	0.856
	10.0	0.100	0.378	0.490	0.931
20	7.0	0.082	0.211	0.250	0.549
	8.0	0.094	0.284	0.381	0.750
	9.0	0.100	0.347	0.460	0.889
	10.0	0.104	0.380	0.510	0.960

4. EFECTO DE ENMASCARAMIENTO EN W

Como se estableció en la Sección 1.2, Bendre y Kale (1985) definieron $M_\lambda = P_1(\lambda) - P_2(\lambda)$ como una medida del efecto de enmascaramiento, dónde

$$P_1(\lambda) = P[W \in A_{n,\alpha} \mid \text{ un outlier está presente }]$$

y

$$P_2(\lambda) = P[W \in A_{n,\alpha} \mid \text{ más de un outlier está presente }].$$

Si la prueba sufre del efecto de enmascaramiento, $P_2(\lambda)$ puede ser menor que $P_1(\lambda)$ e inclusive se puede reducir a cero. Para investigar el comportamiento de W con

respecto al efecto de enmascaramiento, se llevó a cabo un estudio de simulación para $\alpha = 0.05$ con 2000 muestras, cada una de tamaño $n = 10(10)50$ para diferentes valores de λ y valores de k hasta 3.

Los resultados, presentados en las Tablas 3 y 4, muestran un valor positivo de M_λ en cada caso, incrementándose cuando λ se aproxima a cero. Por consiguiente, si dos o más outliers están presentes, la prueba es pobre debido al problema de enmascaramiento.

Tabla 3 Valores de M_λ para $k = 2$ con $T = 9$ y $\alpha = 0.05$

Tamaño de Muestra	λ			
n	0.5	0.2	0.1	0.01
10	0.049	0.086	0.207	0.685
20	0.056	0.026	0.083	0.476
30	0.067	0.014	0.003	0.299
40	0.056	0.016	-0.006	0.219
50	0.104	0.071	0.024	0.184

Tabla 4 Valores de M_λ para $k = 3$ con $T = 9$ y $\alpha = 0.05$

Tamaño de Muestra	λ			
n	0.5	0.2	0.1	0.01
10	0.050	0.144	0.358	0.749
20	0.050	0.050	0.265	0.808
30	0.052	0.001	0.139	0.762
40	0.047	-0.023	0.062	0.662
50	0.088	0.020	0.075	0.598

5. CONCLUSIONES

La presencia de outliers en una muestra siempre ha creado serios problemas. Uno de ellos es por ejemplo la estimación de los parámetros. Cuando las muestras son censuradas se presenta el mismo tipo de inconvenientes. La estadística W se ha

propuesto para detectar un outlier superior en una muestra exponencial censurada. Si se establece que los outliers pueden ocurrir en cualquier lugar de la muestra, se observa que W es potente para valores grandes de T, pero la potencia de la prueba decrece para valores pequeños de T. Además, W está seriamente afectada por el efecto de enmascaramiento. Es decir, que si realmente hay varios outliers presentes en la muestra, W puede fallar en identificar un solo outlier.

BIBLIOGRAFIA

Barnett, V., and Lewis, T. (1984), *Outliers in statistical data*, John Wiley and Sons.

Bartholomew, D. J. (1963), *"The sampling distribution of an estimate arising in life testing "*, Technometrics **5**, 361-374.

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

Bendre, S. M., and Kale, B. K. (1985), *"Masking effect on tests for outliers in exponential models"*, J. Am. Statist. Ass. **80**, 1020-1025.

Gather, U., and Kale, B. K. (1981), *"UMP tests for r-upper outliers in samples from exponential families"*, Instatistics: Applications and new directions, Proc. Indian Stat. Inst. Golden Jubilee Int. Conference, 270-278., Ed. J. K. Ghosh,, Calcutta: Indian Statist. Inst..

Kimber, A. C. (1990), *"Exploratory data analysis for possibly censored data from skewed distributions"*, Appl. Statist. **39**, 21-30.

Lawless, J. F. (1982), *Statistical models and methods for lifetime data*, John Wiley and Sons..

Vargas, J. A., and Iglewicz, B. (1992), *"Detection of one upper outlier in a censored exponential sample"*, Submitted for publication..

Zacks, S. (1986), *"Estimating the scale parameter of an exponential distribution from a sample of time-censored rth-order statistics"*, J. Am. Statist. Ass. **81**, 205-209.