Revista Colombiana de Estadística
Nº 13 Julio de 1986

ESCOGENCIA DE LOS k POSIBLES VALORES EXTREMOS

Lina Sánchez T.
Profesora Asistente
Universidad Nacional

Resumen. Ante la presencia de valores extremos en una muestra, es necesario aplicar criterios objetivos para determinar si esos valores son significativamente diferentes al resto de datos. Se muestra en este artículo cómo la inadecuada escogencia del número de posibles valores extremos, puede hacer que los criterios utilizados detecten más, o menos, valores extremos de los que realmente contiene la muestra. También se presentan algunas alternativas para determinar correctamente el número de valores a examinar.

Abstract. Given the existence of outliers in a sample, the application of objetive criteria is necessary in order to ascertain whether or not these observations are significantly different from the rest of the data. This article shows how an innappropriate initial choice of the number of possible outliers, can force the criteria used to detect more, or in others cases, fewer outliers than the sample really contains, and in addition alternatives are presented in order that the number of outliers may be correctly determined.

1. Introducción.

En la realización de investigaciones sobre diversas áreas es posible encontrar en la muestra observada algún resultado que es diferente a los demás. La presencia de tales valores "extraños" debe llevar a aplicar algún criterio objetivo para establecer si ellos son significativamente diferentes al resto de datos.

Cualquier método para detectar valores extremos requiere de la opinión del investigador, quien al obtener su muestra ha de obser-

var si hay valores "extraños" con relación a los otros o no los hay. Aún cuando el marco teórico que sustenta el estudio, puede dar bases para juzgar si un valor es sorprendente o no, el problema será establecer cuantos valores deben ser tratados como sospechosos y han de ser probados para detectar si son extremos realmente.

La elección del número de valores sospechosos incide en el desempeño de la prueba que se aplica para determinar si hay valores sorprendentes en una muestra. Así, si se toma un número de valores sospechosos más grande que el número de valores extremos que realmente están presentes, la prueba puede declarar como extremos a más valores de los que verdaderamente hay en la muestra, o por el contrario, si se subestima el número de valores extremos presentes, la prueba no detectará todos los valores extremos contenidos en las observaciones.

2. Definiciones Preliminares.

2.1. Valor Extremo.

Es una observación con un residual anor-

mente grande (Anscombe 1960). Esto es, se encuentra localizada lejos del resto de las observaciones en estudio, de tal forma, que a los ojos del investigador es dudosa o sorprendente.

Puede ser generada por errores de ejecución o de medida.

Los valores extremos pueden representarse por modelos que muestran el caso en que las observaciones provienen de dos poblaciones diferentes: la distribución $\eta(\mu,\sigma^2)$, la cual genera los datos "buenos" y la distribución $\eta(\mu+\lambda,\sigma^2)$, o en otro caso, la distribución $\eta(\mu,\lambda^2\sigma^2)$, con $\lambda > 1$, las cuales generan los valores extraños o contaminados.

2.2. Valor Sospechoso.

Es un valor que parece extraño o sorprendente como resultado de una investigación, pero al cual no se le llamará extremo, hasta no aplicar una prueba estadística para establecer si realmente lo es.

Estos valores pueden estar situados en el lado izquierdo de la muestra (m valores), o

al lado derecho (r valores), o a ambos lados. Así $m+r = k$ valores sospechosos en total en la muestra.

En la aplicación de pruebas para detectar valores extremos, generalmente se elige el número de valores sospechosos k y luego se aplica la prueba para $k = 1$, ó $k = 2,...,$ etc, según se ha determinado, actuando así como si k fuera un valor fijo y por el contrario k es una variable aleatoria.

2.3. Potencia de una prueba que detecta valores extremos.

Es la habilidad que tiene el criterio estadístico aplicado, para detectar los valores extremos que verdaderamente tiene la muestra en estudio. Esta habilidad puede medirse según David y Paulson (1965) como sigue:

-La probabilidad de que la prueba concluya correctamente que hay un valor extremo y lo identifique correctamente.

-La probabilidad de que la prueba concluya que hay un valor extremo en la muestra, sin tener en cuenta si identifica o no el va-

lor correcto.

-La probabilidad de que la prueba declare un valor extremo dado que el valor contaminado es el extremo probado.

3. Efectos de la escogencia de k sobre algunos criterios para detectar valores extremos.

Puede darse el caso en el que ante la presencia de un solo valor extremo, si se escoge $k = 2$, es decir, se tratan dos observaciones como sospechosas y se aplica la prueba estadística, ésta declara a ambos valores como extremos.

En este caso, la potencia del criterio usado, se afecta, pues se aumenta el número de valores "falsos positivos" o sea, se declaran más valores extremos de los que realmente hay y ello debido a la inadecuada escogencia de k. Considérense dos pruebas estadísticas en las que se observa este fenómeno:

3.1. Criterio de Rosner.

Se define la estadística de Rosner(1977)

como:

$$R_k = \max_{i \in I_{k-1}} |X_i - a| / b = |X^{(k-1)} - a| / b$$

donde:

$a = \frac{\Sigma X_i}{n-2k}$ es la media de la muestra recortada,

$b = \frac{\Sigma (X_i - a)^2}{n-2k-1}$ es la desviación estándar de la muestra recortada

$I_q = I_0 - (X^{(0)}, X^{(1)}, \ldots, X^{(q-1)})$; $q = 2, .., k-1$

$I_0 = (X_1, X_2, \ldots, X_n)$

La muestra recortada se obtiene quitando el número de valores sospechosos k, de la muestra entera; así, los k valores más grandes y los k valores más pequeños de la muestra ordenada, no se incluyen en el cálculo de a y b.

Un problema potencial al aplicar este criterio es que si se subestima el máximo número de valores extremos presentes en una muestra y se fija en k valores teniendo una muestra con más de k valores extremos, puede ser que la prueba no detecte ningún valor extremo. Esto se debe a que a y particularmente b, pueden ser distorsionadas si en la muestra recortada: $X_{k+1}, X_{k+2}, \ldots, X_{n-k}$, aún hay valores extremos.

Supóngase que en la muestra siguiente hay 1 solo valor extremo:

-3.143	-2.666	-1.305	-0.8980	-0.8138	-0.8138
-0.7577	-0.7437	-0.4771	-0.3087	-0.2526	-0.0982
-0.0842	-0.0561	0.0281	0.1263	0.1684	0.1964
0.2245	0.2947	0.3929	0.4069	0.4209	0.4350
0.4630	0.5472	0.6595	0.7437	1.0800	2.1470

Aplicando el criterio de Rosner el valor de a es -0.1102 y el de b = 0.7490, una vez se han omitido las observaciones -3.143 y 2.147.

$$R_1 = \text{máx}|X_i - a|/b = |-3.143-(-0.1102)|/0.7490 = 4.04$$

La fractila correspondiente, a un nivel de significancia del 5% es:

$R_{1,\alpha}$ = 4.62 (Tabla 1).

Como $R_1 < R_{1,\alpha}$, se concluye que -3.143 no es un valor extremo; es mas se concluye que no hay valores extremos en la muestra.

Sinembargo, si tomamos k = 3, la media recortada es a = -0.037 con b = 0.478

$$R_1 = \max_{i=2,\ldots,30} |X_i - a|/b = |-3.143+0.037|/0.478 = 6.5$$

$$R_2 = \max_{i \in I_1} |X_i - a|/b = |-2.666+0.037|/0.478 = 5.5$$

$$R_3 = \max_{i \in I_2} |X_i - a|/b = |2.147+0.037|/0.478 = 4.51$$

Los valores críticos (tabla 2) a un nivel del 5% son 5.6, 4.32, 3.62 respectivamente, de donde se concluye que -3.143, -2.666, y 2.147 son valores extremos. Se ilustra así el caso en el que la subestimación del valor de k afecta el desempeño del criterio de Rosner.

3.2. Criterio de Tietjen y Moore.

Este criterio para detectar extremos ya sea inferiores o superiores en una muestra proveniente de una población normal con media μ y varianza σ^2, se define como la expresión:

$$E_k = \frac{\Sigma(Z_i - \bar{Z}_k)^2}{\Sigma(Z_i - \bar{Z})^2}$$

donde:

$\bar{Z}$ es la media de la muestra completa.

$\bar{Z}_k$ es la media de las $n-k$ observaciones no ex-

tremas.

Z_i es la X_i cuyo valor absoluto del residual, es el residual más grande; esto es Z_1 será la observación más cercana a $\bar{Z}$ y Z_n será la observación más lejana respecto a $\bar{Z}$.

Al aplicar la estadística E_k, es importante la elección correcta del número de valores sospechosos. Si hay menos valores extremos que los k valores que se van a examinar, puede ser que se rechacen más valores como extremos de los que realmente hay; en el caso contrario, si k es menor que el número de valores extremos que están presentes en la muestra, se puede llegar a concluir que no hay valores extremos.

Considérese la muestra de valores: 399.83, 400.05, 400.71, 401.34, 402.44, 402.47, 402.70, 446.09, en la cual se sospecha que hay dos valores extremos, esto es $k = 2$.

Al aplicar E_k, se tiene: $Z_k = 401.61$ es la media de los X_i omitiendo los dos valores con residuales más grandes.

$$E_2 = \frac{\Sigma(Z_i-\bar{Z}_k)^2}{\Sigma(Z_i-\bar{Z})^2} = 0.03780$$

La fractila correspondiente a un nivel de significancia del 5% (Tabla 3) es $E_{k,\alpha} = 0.099$ Dado que $E_k < E_{k,\alpha}$, se concluye que hay dos valores extremos; aquellos con residuales más grandes: $X_n = 446.09$ y $X_1 = 399.83$. Claramente 446.09 es un valor extremo; pero 399.83 no lo es.

Se han declarado más valores extremos de los que realmente existen, debido a la sobreestimación del valor de k.

4. Alternativas para la escogencia del valor de k.

En el caso de la estadística de Rosner, dado que las medidas a y b como se han definido en 3.1, son susceptibles a la presencia de valores extremos en la muestra recortada, la solución propuesta es usar una medida de tendencia central y una de dispersión, que sean menos sensibles a este hecho; Rosner propone usar la mediana como valor de a y el rango intercuartílico como valor de b.

Tietjen y Moore proponen una forma de escoger el número de valores a probar, para el caso de valores extremos presentes a un solo

lado de la muestra, por ejmplo al lado inferior de la muestra ordenada de menor a mayor. Esta regla consiste en elegir la máxima desviación, respecto a la media, de las variables que están a la izquierda de la media. Se define como el número de observaciones a la izquierda de esa máxima desviación. Luego se aplica el criterio L_k, definido como:

$$L_k = \frac{\Sigma(y_i - \bar{y}_k)^2}{\Sigma(y_i - \bar{y})^2}$$

donde $\bar{y}_k$ es la media de la muestra, después de haber omitido los k valores más pequeños de la muestra.

Autores de otros criterios para detectar valores extremos, tratan de salvar los problemas de la elección de k, con la aplicación sucesiva de sus pruebas; suponen $k = 1$ en cada paso de su procedimiento secuencial y examinan un número de valores sospechosos menor que $n/2$ Si se aplica la prueba a la muestra y se encuentra un valor extremo, este se rechaza y se vuelve a examinar la muestra restante, hasta no encontrar valores extremos.

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., (1983). "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.

David, H.A. and Paulson, A.S., (1965). "The Performance of Several Test for Outliers" Biometrika, 52.

Hawkins, D.M., (1980). *Identifications of Outliers*. Printig House University, Cambridge.

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

McCulloch, C. and Meeter, D. (1983). "Discussión". Technometrics, Vol. 25 #2.

* *

TABLA 1. R_k

Valores Críticos para la Estadística de Rosner para $k = 1, 2$

	α	.10	.05	.01
n	k			
10	1	7.35	8.90	13.38
	2	4.92	5.92	9.13
15	1	5.28	6.01	8.10
	2	3.84	4.31	5.39
20	1	4.64	5.18	6.47
	2	3.50	3.81	4.70
30	1	4.26	4.62	5.51
	2	3.31	3.57	4.15
40	1	4.04	4.41	5.26
	2	3.23	3.43	3.92
50	1	3.98	4.25	4.98
	2	3.20	3.39	3.80
75	1	3.89	4.16	4.77
	2	3.19	3.37	3.72
100	1	3.83	4.09	4.66
	2	3.20	3.34	3.74

TABLA 2. R_k

Valores Críticos
para la Estadística de Rosner
para $k = 1, 2, 3$

	α	.10	.05	.01
n	k			
20	1	5.91	6.60	8.19
	2	4.50	5.06	6.34
	3	3.73	4.16	5.22
30	1	5.07	5.60	6.88
	2	3.93	4.32	5.09
	3	3.35	3.62	4.27
40	1	4.60	5.06	6.05
	2	3.68	3.92	4.53
	3	3.20	3.41	3.82
50	1	4.43	4.76	5.68
	2	3.60	3.82	4.55
	3	3.14	3.30	3.77
75	1	4.18	4.46	5.10
	2	3.47	3.67	4.10
	3	3.08	3.19	3.57
100	1	4.12	4.37	4.98
	2	3.44	3.60	3.88
	3	3.10	3.21	3.45

TABLA 3. E_k

Valores Críticos

para la Estadística de Tietjen y Moore

$\alpha = .05$

n \ k	1	2	3	4	5	6	7	8	9	10
3	.001									
4	.025	.001								
5	.081	.010								
6	.146	.034	.004							
7	.208	.065	.016							
8	.265	.099	.034	.010						
9	.314	.137	.057	.021						
10	.356	.172	.083	.037	.014					
11	.386	.204	.107	.055	.026					
12	.425	.234	.133	.073	.039	.018				
13	.455	.262	.156	.092	.053	.028				
14	.484	.293	.179	.112	.068	.039	.021			
15	.509	.317	.206	.134	.084	.052	.030			
16	.526	.340	.227	.153	.102	.067	.041	.024		
17	.544	.362	.248	.170	.116	.078	.050	.032		
18	.562	.382	.267	.187	.132	.091	.062	.041	.016	
19	.581	.398	.287	.203	.146	.105	.074	.050	.033	
20	.597	.416	.302	.221	.163	.119	.085	.059	.041	.028
25	.652	.493	.381	.298	.236	.187	.146	.114	.089	.068
30	.698	.549	.443	.364	.298	.246	.203	.166	.137	.112
35	.732	.596	.495	.417	.351	.298	.254	.214	.181	.154
40	.758	.629	.534	.458	.395	.343	.297	.259	.223	.195
45	.778	.658	.567	.492	.433	.381	.337	.299	.263	.233
50	.797	.684	.599	.529	.468	.417	.373	.334	.299	.268