Método de “Pruebas Múltiples para Valores Desviados” en el Manejo de Datos Experimentales: Aplicación en Ciencias e Ingenierías

Surendra P. **Verma**

Centro de Investigación en Energía, Universidad Nacional Autónoma de México, Priv. Xochicalco s/no., Col Centro, Apartado Postal 34, Temixco 62580, MEXICO

Tel: (55)56229745; E-mail: spv@cie.unam.mx

Para el manejo de datos experimentales, se dispone de dos grandes tipos de métodos estadísticos: (i) Robustos o de acomodación, debido a que se les considera como robustos contra la presencia de valores desviados; y propiamente (ii) Pruebas múltiples de valores desviados, basadas en la aplicación de una serie de pruebas de discordancia [*Verma 2005; para detalles sobre este libro ver sitio www.cie.unam.mx*]. De acuerdo con el título de esta conferencia plenaria, no se abundará más sobre los métodos robustos, los cuales, probablemente, serán el tema de otra plática.

Los datos experimentales de una determinada variable, por lo general, tienen una distribución normal. La aplicación del método de “**Pruebas múltiples para valores desviados**” en el manejo de datos experimentales se basa en un gran número de pruebas de discordancia, resumidas en la Tabla 1. La metodología básicamente consiste en: (a) aplicar todas las pruebas (Tabla 1) o las seleccionadas bajo cierto criterio, a un conjunto de datos de una variable “normal” física, química, geológica, o de alguna otra naturaleza, con el objeto de determinar los valores discordantes; (b) eliminar estos valores desviados; (c) repetir estos dos procesos (a y b) hasta no encontrar más valores discordantes o, en otras palabras, hasta que los datos representen fielmente una muestra “normal”; y (d) utilizar este conjunto de datos restantes para inferir parámetros estadísticos de tendencia central o de localización (por ejemplo, la media) y de dispersión o de escala (por ejemplo, la desviación estándar), así como para aplicar otras pruebas estadísticas como podrían ser las pruebas F y t de *Student*. Debo enfatizar que el cálculo de este tipo de parámetros como la media y la desviación estándar es válido ***solamente*** si los datos no tienen “*outliers*”, como los obtenidos después de aplicar este procedimiento estadístico.

Tabla 1. Pruebas de discordancia para muestras univariadas normales (modificado de Verma, 2005)

Tipo de Estadístico	Clave prueba	Descripción prueba	Valor(es) probado(s)	Estadístico de la prueba	Significado de la prueba	Prueba n_{min} - n_{max}
Estadístico de Desviación o Dispersión	N1	Más alto	$x_{(n)}$	$TN1_{(u)} = (x_{(n)} - \bar{x}) / s$	Mayor que	3 - 147
		Más bajo	$x_{(1)}$	$TN1_{(l)} = (\bar{x} - x_{(1)}) / s$	Mayor que	3 - 147
	N2	Extremo	$x_{(n)}$ o $x_{(1)}$	$TN2 = Max:\{(x_{(n)} - \bar{x}) / s, (\bar{x} - x_{(1)}) / s\}$	Mayor que	3 - 20
	N3	k=2 Más alto	$x_{(n)}, x_{(n-1)}$	$TN3_{(2u)} = (x_{(n)} + x_{(n-1)} - 2\bar{x}) / s$	Mayor que	5 - 100
		k=3 Más alto	$x_{(n)}, x_{(n-1)}, x_{(n-2)}$	$TN3_{(3u)} = (x_{(n)} + x_{(n-1)} + x_{(n-2)} - 3\bar{x}) / s$	Mayor que	7 - 100
		k=4 Más alto	$x_{(n)}, x_{(n-1)}, x_{(n-2)}, x_{(n-3)}$	$TN3_{(4u)} = (x_{(n)} + x_{(n-1)} + x_{(n-2)} + x_{(n-3)} - 4\bar{x}) / s$	Mayor que	9 - 100
		k=2 Más bajo	$x_{(1)}, x_{(2)}$	$TN3_{(2l)} = (2\bar{x} - x_{(1)} - x_{(2)}) / s$	Mayor que	5 - 100
		k=3 Más bajo	$x_{(1)}, x_{(2)}, x_{(3)}$	$TN3_{(3l)} = (3\bar{x} - x_{(1)} - x_{(2)} - x_{(3)}) / s$	Mayor que	7 -100
		k=4 Más bajo	$x_{(1)}, x_{(2)}, x_{(3),} x_{(4)}$	$TN3_{(4l)} = (4\bar{x} - x_{(1)} - x_{(2)} - x_{(3)} - x_{(4)}) / s$	Mayor que	9 - 100
Estadístico de Sumas de Cuadrados	N4	k=1 Más alto	$x_{(n)}$	$TN4_{(1u)} = S^2_{(n)} / S^2$	Menor que	3 - 50
		k=2 Más alto	$x_{(n)}, x_{(n-1)}$	$TN4_{(2u)} = S^2_{(n),(n-1)} / S^2$	Menor que	4 - 149
		k=3 Más alto	$x_{(n)}, x_{(n-1)}, x_{(n-2)}$	$TN4_{(3u)} = S^2_{(n),(n-1),(n-2)} / S^2$	Menor que	6 - 50
		k=4 Más alto	$x_{(n)}, x_{(n-1)}, x_{(n-2)}, x_{(n-3)}$	$TN4_{(4u)} = S^2_{(n),(n-1),(n-2),(n-3)} / S^2$	Menor que	8 - 50

Tabla 1 (continuación - 1). Pruebas de discordancia para muestras univariadas normales (modificado de Verma, 2005)

Tipo de Estadístico	Clave prueba	Descripción prueba	Valor(es) probado(s)	Estadístico de la prueba	Significado de la prueba	Prueba n_{min} - n_{max}
Estadístico de Sumas de Cuadrados	N4	K=1 Más bajo	$x_{(1)}$	$TN4_{(1l)} = S^2_{(1)} / S^2$	Menor que	3 - 50
	(sigue)	K=2 Más bajo	$x_{(1)}, x_{(2)}$	$TN4_{(l2)} = S^2_{(1),(2)} / S^2$	Menor que	4 - 149
		K=3 Más bajos	$x_{(1)}, x_{(2)}, x_{(3)}$	$TN4_{(3l)} = S^2_{(1),(2),(3)} / S^2$	Menor que	6 - 50
		K=4 Más bajos	$x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}$	$TN4_{(4l)} = S^2_{(1),(2),(3),(4)} / S^2$	Menor que	8 - 50
(Continuación)	N5	K=2 Más alto – más bajo	$x_{(n)}, x_{(1)}$	$TN5_{(ul)} = S^2_{(n),(1)} / S^2$	Menor que	4 - 100
Estadístico Intervalo total/ Dispersión	N6	Más alto – más bajo	$x_{(n)}, x_{(1)}$	$TN6_{(ul)} = (x_{(n)} - x_{(1)}) / s$	Mayor que	3 -1000
Estadístico de Exceso / Dispersión	N7	Más alto	$x_{(n)}$	$TN7_{(u)} = (x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(1)})$	Mayor que	3 - 30
	N8	Extremo	$x_{(n)}$ o $x_{(1)}$	$TN8 = Max: \{(x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(1)}), (x_{(2)} - x_{(1)}) / (x_{(n)} - x_{(1)})\}$	Mayor que	4 - 30
	N9	Más alto	$x_{(n)}$	$TN9_{(u)} = (x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(2)})$	Mayor que	4 - 30
		Más bajo	$x_{(1)}$	$TN9_{(l)} = (x_{(2)} - x_{(1)}) / (x_{(n-1)} - x_{(1)})$	Mayor que	4 - 30
	N10	Más alto	$x_{(n)}$	$TN10_{(u)} = (x_{(n)} - x_{(n-1)}) / (x_{(n)} - x_{(3)})$	Mayor que	5 - 30
		Más bajo	$x_{(1)}$	$TN10_{(l)} = (x_{(2)} - x_{(1)}) / (x_{(n-2)} - x_{(1)})$	Mayor que	5 - 30

Tabla 1 (continuación - 2). Pruebas de discordancia para muestras univariadas normales (modificado de Verma, 2005)

Tipo de Estadístico	Clave prueba	Descripción prueba	Valor(es) probado(s)	Estadístico de la prueba	Significado de la prueba	Prueba n_{min} - n_{max}
Estadístico de Exceso / Dispersión (continuación)	N11	Dos valores más altos	$x_{(n)}, x_{(n-1)}$	$TN11_{(2u)} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(1)})$	Mayor que	4 - 30
		Dos valores más bajos	$x_{(1)}, x_{(2)}$	$TN11_{(2l)} = (x_{(3)} - x_{(1)})/(x_{(n)} - x_{(1)})$	Mayor que	4 - 30
	N12	Dos valores más altos	$x_{(n)}, x_{(n-1)}$	$TN12_{(2u)} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(2)})$	Mayor que	5 - 30
		Dos valores más bajos	$x_{(1)}, x_{(2)}$	$TN12_{(2l)} = (x_{(3)} - x_{(1)})/(x_{(n-1)} - x_{(1)})$	Mayor que	5 - 30
	N13	Dos valores más altos	$x_{(n)}, x_{(n-1)}$	$TN13_{(2u)} = (x_{(n)} - x_{(n-2)})/(x_{(n)} - x_{(3)})$	Mayor que	6 - 30
		Dos valores más bajos	$x_{(1)}, x_{(2)}$	$TN13_{(2l)} = (x_{(3)} - x_{(1)})/(x_{(n-2)} - x_{(1)})$	Mayor que	6 - 30
Estadístico de Momento De alto orden	N14	Extremo	$x_{(n)}$ o $x_{(1)}$	$TN14 = \left[\frac{n^{1/2}\{\sum_{i=1}^{n}(x_i - \bar{x})^3\}}{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^{3/2}} \right]$	Mayor que	5 - 1000
	N15	Extremo	$x_{(n)}$ o $x_{(1)}$	$TN15 = \left[\frac{n\{\sum_{i=1}^{n}(x_i - \bar{x})^4\}}{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^{2}} \right]$	Mayor que	5 - 1000

Para mayores detalles sobre la simbología usada en esta tabla, ver el libro [*Verma 2005*; páginas 95-97, en particular y Capítulo 4, en general].

La manera de aplicar una determinada prueba es que, en un conjunto de datos, se identifica(n) el o los posible(s) valor(es) a probar, se calcula el estadístico de la prueba y al valor obtenido de este estadístico se le compara con el valor crítico para la determinada prueba y un determinado nivel de confianza (99% según [*Verma, 2005*]). Si el valor calculado es mayor o menor que (dependiendo de la naturaleza de la

prueba; ver Tabla 1), se determina que el o los valor(es) probado(s) es(son) un dato(s) discordante(s) y debe(n) ser eliminado(s). Se aplican todas las pruebas seleccionadas y se toma la decisión de eliminar los valores desviados hasta que ninguna de las pruebas seleccionadas señale que los datos restantes tienen valores desviados. Después de aplicar este procedimiento, los datos finales representarían una muestra estadística de una determinada población normal, sin ninguna "contaminación", y los "contaminantes" tendrían cierta utilidad para la interpretación de los datos.

Un hecho importante a mencionar es que los valores críticos disponibles en la literatura son poco precisos y exactos, además de ser limitados en el número máximo de datos a los que una determinada prueba pueda aplicarse (ver el valor de n_{max} en la última columna de la Tabla 1 y los libros [*Barnett y Lewis 1994*; *Verma 2005*]). Además de esto, aún dentro de los límites (n_{min} - n_{max}) no se dispone de los valores críticos para muchos casos. Por ejemplo, para la prueba N2, se conocen los valores críticos (con una precisión de solamente dos puntos decimales) sólo para n = 3, 4, 5, 6, 7, 8, 9, 10, 12, 15 y 20 [*Verma 2005*]. Nótese la falta de los valores críticos para n = 11, 13, 14, 16, 17, 18 y 19, los cuales necesitan obtenerse por algún tipo de interpolación. De igual manera, para las ampliamente usadas pruebas de Dixon (N7, N9-N13; ver Tabla 1), los valores críticos se disponen para n hasta 30. Esto limita seriamente la aplicación de muchas de las pruebas para datos experimentales.

Reconociendo estas limitaciones, se ha desarrollando, recientemente, una metodología de simulación tipo *Monte Carlo*, muy precisa y exacta, y se han obtenido nuevos valores críticos para hasta 100 datos y para todas las pruebas de Dixon [*Verma y Quiroz-Ruiz 2006a*] y para las 9 pruebas restantes con sus 22 variantes resumidas en la Tabla 1 [*Verma y Quiroz-Ruiz 2006b*].

El presente método de "***Pruebas Múltiples para Valores Desviados***" con nuevos valores críticos [*Verma y Quiroz-Ruiz 2006a, b*] para un total de 15 pruebas y 33 variantes y para hasta 100 datos proporciona mejores resultados que el método de la gráfica de "***Box y Whisker***" usado por algunos investigadores europeos. Por otra parte, el así llamado método de "***Dos desviaciones estándar***" que ha sido usado para procesar las bases de datos "entre-laboratorios" por investigadores

norteamericanos y japoneses, ha sido demostrado erróneo, en múltiples ocasiones, por Verma y sus colaboradores, y por lo tanto, debe ser abandonado.

En su lugar, propongo que debe usarse el presente método de "***Pruebas Múltiples***" con, al menos, 15 pruebas y 33 variantes, todas ellas ahora aplicables para los tamaños de muestras hasta 100. Para muestras más grandes (>100), aunque sí se pueden aplicar algunas de las pruebas, debo informarles que se encuentra en proceso un nuevo trabajo de investigación para simular valores críticos, más precisos y exactos que los disponibles hasta ahora sólo para algunas de estas pruebas.

Es importante mencionar que se han logrado aplicaciones de presente método de "***Pruebas Múltiples para valores desviados***" en una gran variedad de campos de Ciencias e Ingenierías, tales como Agricultura, Astronomía, Biología, Biomedicina, Biotecnología, Ciencia del Suelo, Ciencia Nuclear, Ciencia y Tecnología de los Alimentos, Contaminación Ambiental, Electrónica, Geocronología, Geología Estructural, Geología Isotópica, Geoquímica, Investigación del Agua, Investigación del Petróleo, Meteorología, Paleontología, Programas de Aseguramiento de Calidad, Química y Zoología. En conclusión, esta metodología de "***Pruebas Múltiples para Valores Desviados***", aplicada y ahora ampliada por el grupo del CIE-UNAM, puede aplicarse en estos y otros campos de conocimiento en Ciencias e Ingenierías.

[nota importante: debido a la falta de espacio – disponibilidad de 6 páginas, se citan aquí pocas referencias; se recomienda al lector examinar el gran número de referencias incluidas en los trabajos que se señalan a continuación.]

Referencias

Barnett, V., Lewis, T., 1994, *Outliers in Statistical Data*. Third edition: Chichester, John Wiley, 584 p.

Verma, S.P., 2005, *Estadística Básica para el Manejo de Datos Experimentales: Aplicación en la Geoquímica (Geoquimiometría)*: México, D. F., Universidad Nacional Autónoma de México, 186 p.

Verma, S.P., Quiroz-Ruiz, A., 2006a, Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. ***Revista Mexicana de Ciencias Geológicas*** 23, 133-161. *Trabajo completo disponible en el sitio de internet:* http://satori.geociencias.unam.mx/

Verma, S.P., Quiroz-Ruiz, A., 2006b, Critical values for 22 discordancy test variants for outliers in normal samples up to sizes 100, and applications in science and engineering. ***Revista Mexicana de Ciencias Geológicas*** (enviado en abril de 2006, se encuentra en evaluación).