PXS
Performance Excellence Solutions

EL ANÁLISIS DE PUNTOS ATÍPICOS

Rolando Guido Sáenz

El Análisis de Puntos Atípicos

Antes de iniciar quisiera pedir disculpas al lector por no utilizar la notación oficial vigente el en país de separar los valores decimales con una coma. En el presente documento se estará utilizando el punto como separador decimal. Este problema lo heredé hace aproximadamente 50 años cuando cursaba mis estudios primarios y nunca me pude sobreponer a las deficiencias de mi formación inicial. Espero que esto no sea un problema fundamental para lo que se quiere exponer aquí. Sin más preámbulo, a lo que interesa.

Un aspecto importante en el análisis de datos es la determinación de puntos atípicos en la muestra. Hay un libro muy profundo sobre el tema escrito por Charu C. Aggarwal llamado "Outlier Analysis" que es una verdadera joya y del que voy a extraer algo de lo que aquí leerán, aunque por la profundidad del texto, y mis limitaciones, será de las partes más simples. El Sr Aggarwal es un indio privilegiado que ha conseguido un PhD del MIT y es miembro distinguido del IBM T.J. Watson Research Center cuya biografía y currículo se consigue en internet y es impresionante. He extraído una buena parte de otro libro de estadística para la confiabilidad de Mark Allen Durivage.

La detección de valores atípicos es importante en una gran cantidad actividades humanas como por ejemplo:

- Detección de fraudes en compras con tarjetas de crédito
- Detección de solicitudes fraudulentas de crédito, o la detección de clientes problemáticos
- Detección de accesos no autorizados en un servidor o en transacciones bursátiles
- Análisis del desempeño de una red de computadoras que lleven a identificar cuellos de botella
- Diagnóstico de fallas en motores, generadores, oleoductos, o satélites
- Detección de fallas estructurales, como pueden ser vigas fracturadas
- Análisis de imágenes de satélite que identifiquen características nuevas o mal identificadas previamente
- Detección de cambios en imágenes de las cámaras instaladas en sistemas de visión inteligentes en sistemas de vigilancia o en robots
- Segmentación del movimiento para identificar movimientos independientes del fondo de la imagen
- Monitoreo de series de tiempo como la información de un electrocardiograma y otras técnicas de diagnóstico médico para la detección de enfermedades
- En la industria farmacéutica, para detectar estructuras celulares nuevas
- En la determinación de las características de calidad de los productos que se manufacturan
- En la medición de la temperatura superficial del océano para identificar patrones anormales del clima, como por ejemplo el fenómeno del Niño

En estas notas solo se considerarán datos unidimensionales, de forma que solo se analizará si uno o varios datos son muy grandes o muy pequeños para ser considerados atípicos. Los métodos que se analizarán serán aplicables a poblaciones normales con excepción del que se basa en el teorema de

Bienayme - Chebyshev (usualmente conocido como el teorema de Chebyshev) que aunque no solo esa existe, es la única que se presenta.

Lo que es un valor atípico

Un valor no es atípico por sí mismo. Los valores serán atípicos en relación con el modelo que describe el comportamiento del fenómeno que se está estudiando. Esto quiere decir que para medir si un valor es atípico se debe conocer la distribución probabilística de los valores, y todo a través de una muestra. No es lo mismo que la distribución sea normal o que sea exponencial. Conocer los parámetros de una distribución siempre es difícil, a menos que la población sea muy pequeña y se tengan todos los individuos, pero en este caso no existen valores atípicos. Así es la población y punto. En la mayoría de los casos se toma una muestra y se asume que viene de una población, el problema es que, primero no se conoce cómo es la población, y segundo se asume que todos los individuos que se seleccionaron vienen de la misma población que no se conoce. Pero como analistas de datos no se tienen muchas opciones. Se tiene una muestra y se analiza, se estiman los parámetros de la población (la que parece más probable) y se decide si algún valor es atípico o no. Aquí la fuerza viene de la muestra.

Las formas gráficas

Una forma gráfica de evaluar valores atípicos puede hacerse utilizando un histograma. Los datos que se muestran en este ejemplo son tomados del libro Engineering Statistics Handbook del NIST y modificados en algunos casos y están en la “Tabla 1. Datos para el cálculo de valores atípicos”.

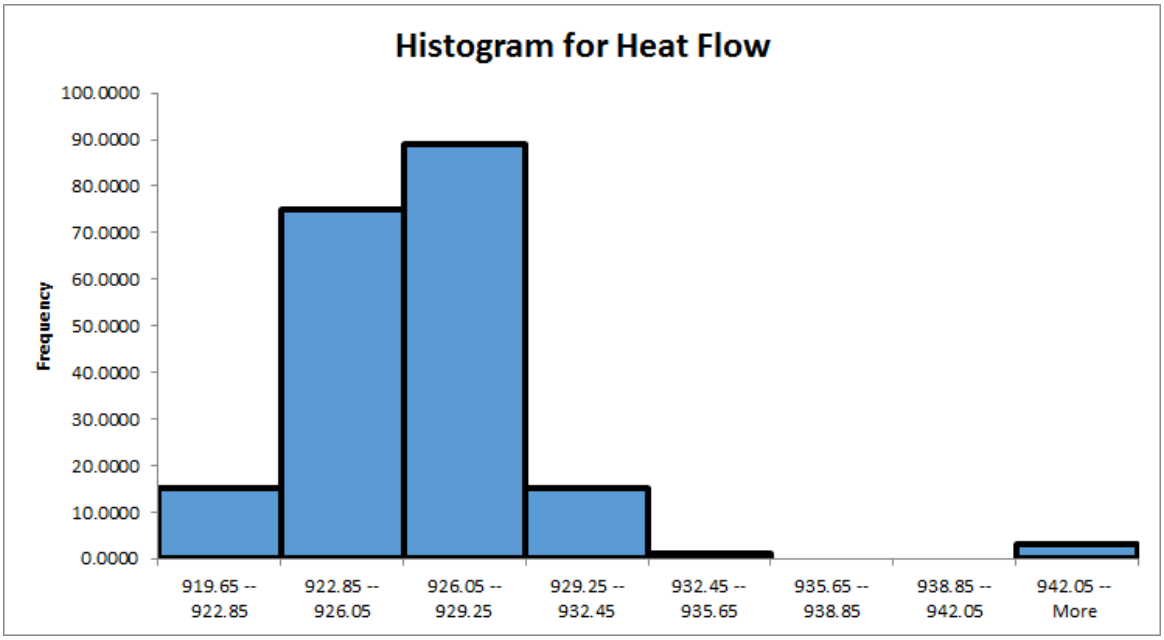

Gráfico 1. Histograma mostrando valores atípicos

El problema con los métodos gráficos es que no determinan cuáles son los valores atípicos y solo se puede adivinar cuáles son, cuando son evidentes.

Otra forma gráfica es utilizar un Diagrama de Caja o Box Plot. El Gráfico 2. Diagrama de caja mostrando valores atípicos muestra los mismos datos que se utilizaron en el histograma, sin embargo note que en la parte inferior del rango se muestra como atípico un valor, que no era evidente en el histograma. Más adelante se probará cuantitativamente si es atípico o no.

A la hora de clasificar un valor como atípico se debe diferenciar tres áreas en el rango de los datos. El primero es el rango de los datos típicos del modelo, el segundo es el rango que el Sr. Aggarwal llama el rango del ruido (atípicos moderados) y el tercero que es rango de los valores atípicos (atípicos extremos). Note que los dos últimos, en el caso de la distribución normal pueden ocurrir en la parte baja o en la parte alta de los valores. En un Diagrama de Caja cualquier valor que se encuentre a más de 2.7 desviaciones estándar de la media se clasificará como un valor atípico moderado siempre que no supere las 4.7 desviaciones estándar. Los valores que superen esta última cifra serán clasificados como valores atípicos extremos. Si calcula la probabilidad de que un valor se encuentre a más y menos 2.7 desviaciones estándar de la media se obtiene un valor de 0.993.

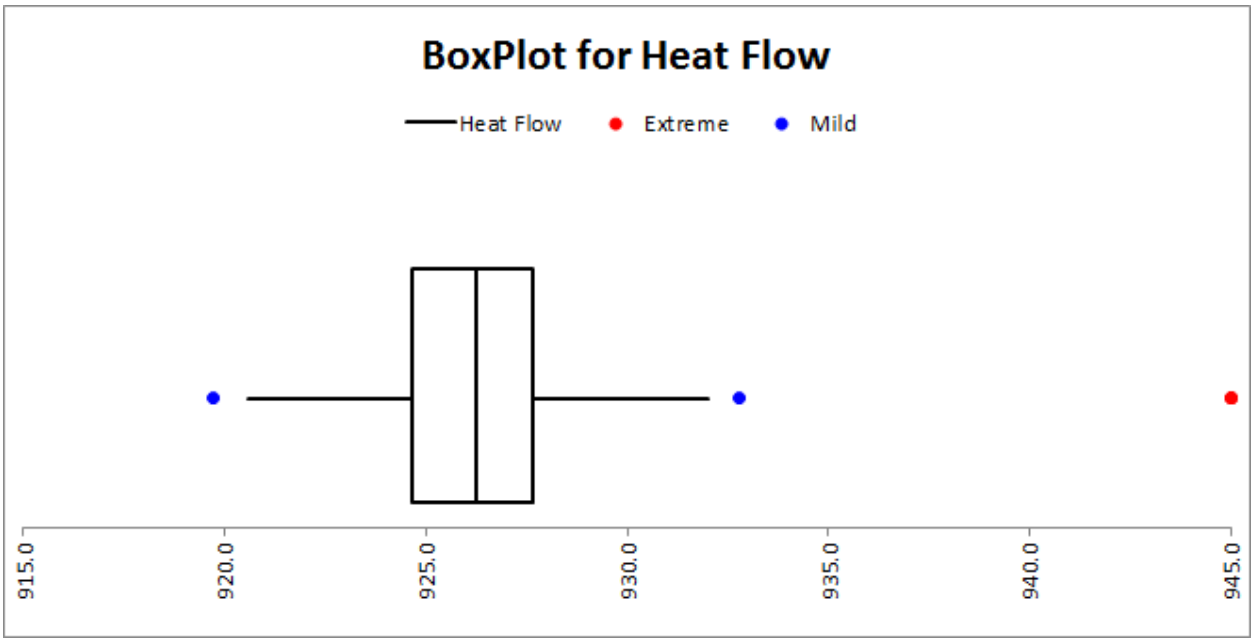

Gráfico 2. Diagrama de caja mostrando valores atípicos

Los valores típicos cubren entonces el 99.3% de la parte central de la distribución normal. Si se calcula la probabilidad para los valores atípicos extremos según el criterio que se muestra en un Diagrama de Caja, de una distribución normal se obtiene un valor de 0.0000023, (0.00023% o una oportunidad en aproximadamente 426100). La probabilidad de los valores atípicos moderados se puede calcular por diferencia con los valores de probabilidad anteriores y es de 0.00697 o 0.697%.

Los métodos cuantitativos

Método 1. Utilizar la desviación estándar de una distribución normal para detectar valores atípicos.

Este método es muy parecido al que utiliza el Diagrama de Caja, la diferencia es que usualmente se consideran valores atípicos los que superan las 3 desviaciones estándar a ambos lados de la media, y no están por lo tanto contemplados entre el 99.73% de los datos centrales de la distribución. Utilizando este método con los datos del ejemplo anterior y calculando los valores a más y menos 3 desviaciones estándar éstos serían 936.1 y 916.7. Una revisión indica que solo tres valores están por encima del límite superior de 936.1 y tienen un valor los tres de 945.

Heat Flow							
920.6	925.5	922.1	924.1	923.8	926.1	925.3	927.8
930.0	926.9	927.1	923.9	926.9	927.4	926.1	928.5
927.8	929.0	925.2	926.4	924.8	929.2	923.8	924.0
930.6	927.4	928.1	924.3	925.7	927.1	922.5	926.8
927.5	925.6	927.1	924.7	926.8	926.7	923.6	924.8
928.9	926.2	929.5	925.2	928.8	930.9	924.0	922.5
928.7	925.0	930.2	926.2	925.8	926.4	926.4	923.1
926.1	926.2	927.9	924.7	928.6	927.9	924.4	927.0
930.3	926.4	923.7	930.6	925.1	925.5	927.8	926.5
927.6	926.5	923.4	923.8	925.7	922.9	931.1	928.4
927.3	924.2	924.5	924.9	926.8	925.3	926.2	928.1
928.8	924.0	922.2	925.7	929.1	925.6	926.0	926.3
925.6	922.2	920.7	926.6	921.9	926.3	925.3	929.2
925.2	924.2	925.9	929.9	927.0	922.0	924.6	925.2
929.8	921.5	927.6	924.5	921.9	925.8	928.4	924.4
926.7	928.6	926.9	928.7	924.1	926.8	925.1	928.3
925.7	927.2	925.7	930.1	927.0	926.8	927.5	919.7
927.8	926.6	926.5	925.7	922.7	924.9	925.5	923.1
924.8	928.5	929.6	927.1	925.9	923.5	928.0	923.3
925.2	926.9	929.3	927.5	928.6	924.3	927.5	923.5
927.6	926.8	926.4	928.2	932.0	925.3	926.2	921.7
927.9	924.6	928.1	925.3	932.8	926.3	927.5	927.4
926.7	923.1	926.7	926.9	926.3	924.3	925.2	927.4
924.6	924.1	930.1	928.2	924.8	926.1	923.0	945.0
945.0	945.0	923.9	926.0	922.5	925.3		

Tabla 1. Datos para el cálculo de valores atípicos

Método 2. Prueba del Valor Discordante

Es parecido al método anterior de la desviación estándar, pero el valor calculado (D) se compara con un valor de tabla.

El valor de D se calcula con la siguiente fórmula:

$$D = \frac{|\mu - x_i|}{\sigma}$$

Dónde:

D: Valor de la Discordancia

μ: Media Aritmética

σ: Desviación Estándar

x_i: Potencial valor atípico

$D_{(\alpha,n)}$: Valor crítico para un nivel de confianza alfa (α) y una muestra de tamaño n.

Como la tabla que se muestra contiene valores críticos para tamaños de muestra hasta de 50 unidades, el análisis se aplicará a los últimos 46 valores de la tabla anterior y los valores críticos se muestran en la tabla siguiente.

Valores críticos para la prueba del Valor Discordante

n	alfa=0.01	alfa 0.05	n	alfa=0.01	alfa 0.05
3	1.155	1.153	27	3.049	2.698
4	1.492	1.463	28	3.068	2.714
5	1.749	1.672	29	3.085	2.73
6	1.944	1.822	30	3.103	2.745
7	2.097	1.938	31	3.119	2.759
8	2.221	2.032	32	3.135	2.773
9	2.323	2.11	33	3.15	2.786
10	2.41	2.176	34	3.164	2.799
11	2.485	2.234	35	3.178	2.811
12	2.55	2.285	36	3.191	2.823
13	2.607	2.331	37	3.204	2.835
14	2.659	2.371	38	3.216	2.846
15	2.705	2.409	39	3.228	2.857
16	2.747	2.443	40	3.24	2.866
17	2.785	2.475	41	3.251	2.877
18	2.821	2.504	42	3.261	2.887
19	2.854	2.532	43	3.271	2.896
20	2.884	2.557	44	3.282	2.905
21	2.912	2.58	45	3.292	2.914
22	2.939	2.603	46	3.302	2.923
23	2.963	2.624	47	3.31	2.931
24	2.987	2.644	48	3.319	2.94
25	3.009	2.663	49	3.329	2.948
26	3.029	2.681	50	3.336	2.956

Tabla 2. Valores críticos del estadístico D

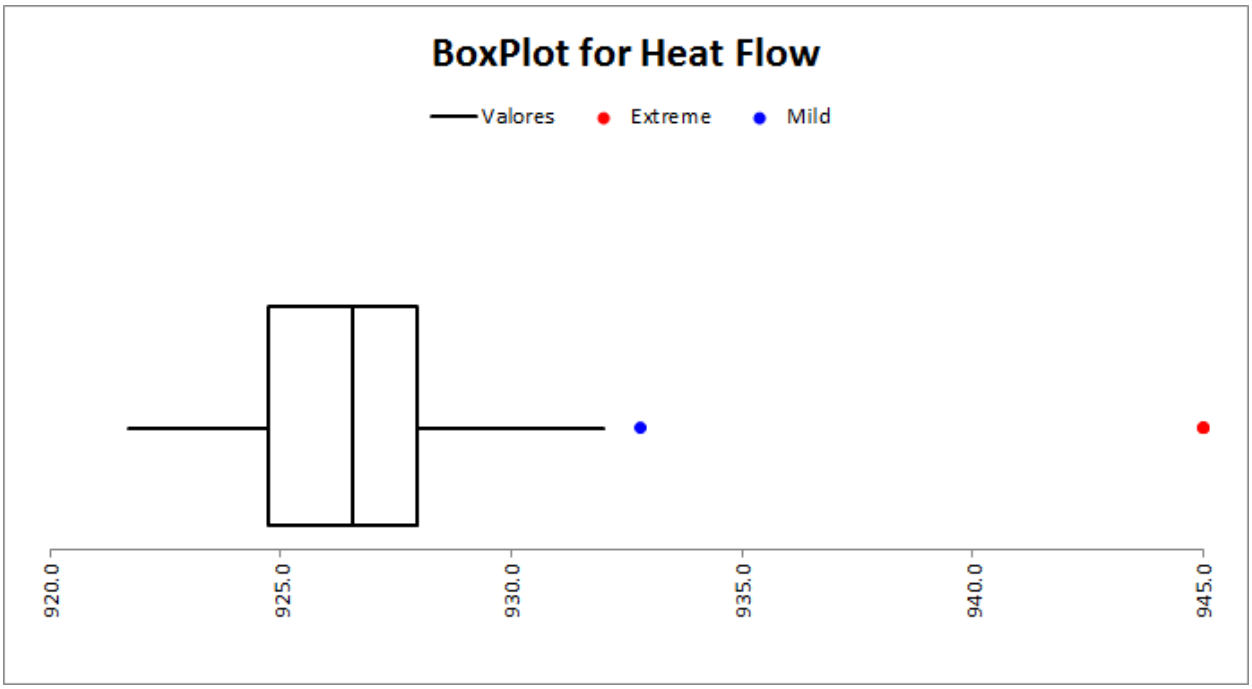

Gráfico 3 Diagrama de Caja de los últimos 46 valores

Se desea calcular si el valor de 923.8 se puede considerar un valor atípico, utilizando la prueba del valor discordante. El valor calculado para D es 1.017. El valor crítico de la tabla para alfa de 0.05 y n de 46 es de 2.923. Como el valor no supera el valor de la tabla no hay evidencia para decir que es un valor atípico. Si se hace el mismo cálculo con el valor de 945 el cálculo del Estadístico D es 3.379, por lo que se debe considerar un valor atípico.

Método 3. Prueba de Nalimov

La prueba es similar a la anterior pero el estadístico q se calcula:

$$q = \frac{|\mu - x_i|}{\sigma} \sqrt{\frac{n}{n-1}}$$

Dónde:

q: Valor del estadístico con n-2 grados de libertad
μ: Media Aritmética
σ: Desviación Estándar
x_i: Potencial valor atípico
n: Tamaño de la muestra

Valores Críticos para la Prueba de Nalimov

gl	qcrit α=0.05	qcrit α=0.01	qcrit α=0.001	gl	qcrit α=0.05	qcrit α=0.01	qcrit α=0.001
1	1.409	1.414	1.414	19	1.936	2.454	2.975
2	1.645	1.715	1.73	20	1.937	2.46	2.99
3	1.757	1.918	1.982	25	1.942	2.483	3.047
4	1.814	2.051	2.178	30	1.945	2.498	3.085
5	1.848	2.142	2.329	35	1.948	2.509	3.113
6	1.87	2.208	2.447	40	1.949	2.518	3.134
7	1.885	2.256	2.54	45	1.95	2.524	3.152
8	1.895	2.294	2.616	50	1.951	2.529	3.166
9	1.903	2.324	2.678	100	1.956	2.553	3.227
10	1.91	2.348	2.73	200	1.958	2.564	3.265
11	1.916	2.368	2.774	300	1.958	2.566	3.271
12	1.92	2.385	2.812	400	1.959	2.568	3.275
13	1.923	2.399	2.845	500	1.959	2.57	3.279
14	1.926	2.412	2.874	600	1.959	2.571	3.281
15	1.928	2.423	2.899	700	1.959	2.572	3.283
16	1.931	2.432	2.921	800	1.959	2.573	3.285
17	1.933	2.44	2.941	1000	1.96	2.576	3.291
18	1.935	2.447	2.959				

Tabla 3 Valores críticos de q

Utilizando los 198 valores originales el valor de q para 932.8 y 945 respectivamente es 1.975 y 5.758. El valor de la tabla para 200 (y note que no hay mucha diferencia si fueran 100 o 1000) y un alfa de 0.05 indica que q crítico es igual a 1.958. Por lo que ambos serían valores atípicos. Sin embargo basado en mi experiencia, esta prueba es muy estricta y tiende a señalar más valores atípicos que las otras pruebas.

Método 4. Extensión Camp Meidell del teorema de Bienayme – Chebyshev

La extensión del teorema de Bienayme – Chebyshev (lo pueden buscar en Wikipedia como Chebyshev's Inequality) especifica que para distribuciones continuas, unimodales y simétricas y para valores de k superiores a 1.155 (2 dividido por la raíz de 3) la probabilidad de que la distancia (valor absoluto de la diferencia) de un valor y el promedio de los valores sea mayor o igual que k veces la desviación estándar es menor o igual que $4/9k^2$. En terminología matemática sería:

$$P(|x_i - \mu| \geq k\sigma) \leq 4/9k^2$$

Dónde:

$k \geq 2/\sqrt{3}$

μ: Media Aritmética

σ: Desviación Estándar

x_i: Potencial valor atípico

Si se evalúa los valores 945 y 932.8 tienen un valor de k de 5.74 y 1.97 respectivamente (la diferencia entre los valores y la media de toda la muestra es de 5.74 sigmas y 1.97 sigmas). La probabilidad de que

se den estos valores es menor o igual que 0.013 y 0.115. Si se considera los valores como si fueran Valores-p, se concluye que el primero es un valor atípico y el segundo no.

Método5. Prueba de Dean y Dixon

La prueba es especial para muestras pequeñas y para utilizarla es necesario que los valores se ordenen de menor a mayor. Las fórmulas que se usan dependen de la cantidad de valores analizados.

	Para el valor más pequeño	Para el valor más grande
Entre 3 y 7 valores:	$r_{10} = \frac{X_2 - X_1}{X_n - X_1}$	$r_{10} = \frac{X_n - X_{n-1}}{X_n - X_1}$
Entre 8 y 10 valores:	$r_{11} = \frac{X_2 - X_1}{X_{n-1} - X_1}$	$r_{11} = \frac{X_n - X_{n-1}}{X_n - X_2}$
Entre 11 y 13 valores:	$r_{21} = \frac{X_3 - X_1}{X_{n-1} - X_1}$	$r_{21} = \frac{X_n - X_{n-2}}{X_n - X_2}$
Para 14 y más valores:	$r_{22} = \frac{X_3 - X_1}{X_{n-2} - X_1}$	$r_{22} = \frac{X_n - X_{n-2}}{X_n - X_3}$

Dónde n es la cantidad de valores

Si por ejemplo se tiene los siguientes valores:

1, 3, 6, 7, 8, 9, 10, 11, 12, 23

N=10

Se desea saber si el 23 es un valor atípico.

El valor de r_{11} para probar el valor más alto = (23-12)/(23-3)=0.55

El valor crítico de r_{11} para n=10 y alfa de 0.05 es: 0.477 por lo que 23 es un valor atípico.

Se desea saber si el 1 es un valor atípico.

El valor de r_{11} para probar el valor más bajo = (3-1)/(12-1)=0.182

El valor es menor que el valor crítico de 0.477, se concluye que 1 no es un valor atípico.

Valores críticos para R_{10}

N	α=0.001	α=0.002	α=0.005	α=0.01	α=0.02	α=0.05	α=0.1	α=0.2
3	0.999	0.998	0.994	0.988	0.976	0.941	0.886	0.782
4	0.964	0.949	0.921	0.889	0.847	0.766	0.679	0.561
5	0.895	0.869	0.824	0.782	0.729	0.643	0.559	0.452
6	0.822	0.792	0.744	0.698	0.646	0.563	0.484	0.387
7	0.763	0.731	0.681	0.636	0.587	0.507	0.433	0.344

Valores críticos para R_{11}

N	α=0.001	α=0.002	α=0.005	α=0.01	α=0.02	α=0.05	α=0.1	α=0.2
8	0.799	0.769	0.724	0.682	0.633	0.554	0.48	0.386
9	0.75	0.72	0.675	0.634	0.586	0.512	0.441	0.352
10	0.713	0.683	0.637	0.597	0.551	0.477	0.409	0.325

Valores críticos para R_{21}

N	α=0.001	α=0.002	α=0.005	α=0.01	α=0.02	α=0.05	α=0.1	α=0.2
11	0.77	0.746	0.708	0.674	0.636	0.575	0.518	0.445
12	0.739	0.714	0.676	0.643	0.605	0.546	0.489	0.42
13	0.713	0.687	0.649	0.617	0.58	0.522	0.467	0.399

Valores críticos para R_{22}

N	α=0.001	α=0.002	α=0.005	α=0.01	α=0.02	α=0.05	α=0.1	α=0.2
14	0.732	0.708	0.672	0.64	0.603	0.546	0.491	0.422
15	0.708	0.685	0.648	0.617	0.582	0.524	0.47	0.403
16	0.691	0.667	0.63	0.598	0.562	0.505	0.453	0.386
17	0.671	0.647	0.611	0.58	0.545	0.489	0.437	0.373
18	0.652	0.628	0.594	0.564	0.529	0.475	0.424	0.361
19	0.64	0.617	0.581	0.551	0.517	0.462	0.412	0.349
20	0.627	0.604	0.568	0.538	0.503	0.45	0.401	0.339
25	0.574	0.55	0.517	0.489	0.457	0.406	0.359	0.302
30	0.539	0.517	0.484	0.456	0.425	0.376	0.332	0.278
35	0.511	0.49	0.459	0.431	0.4	0.354	0.311	0.26
40	0.49	0.469	0.438	0.412	0.382	0.337	0.295	0.246
45	0.475	0.454	0.423	0.397	0.368	0.323	0.283	0.234
50	0.46	0.439	0.41	0.384	0.355	0.312	0.272	0.226
60	0.437	0.417	0.388	0.363	0.336	0.294	0.256	0.211
70	0.422	0.403	0.374	0.349	0.321	0.28	0.244	0.201
80	0.408	0.389	0.36	0.337	0.31	0.27	0.234	0.192
90	0.397	0.377	0.35	0.326	0.3	0.261	0.226	0.185
100	0.387	0.368	0.341	0.317	0.292	0.253	0.219	0.179

Tabla 4. Valores críticos para la prueba de valores atípicos Dean y Dixon.

Método 6. Prueba de Grubbs

La prueba se utiliza para probar los valores máximo o mínimo de una serie de datos. Se calcula un estadístico g según sea el valor mínimo o el máximo. Si un valor resulta atípico y se desea probar si otro

también lo es, se debe eliminar el atípico y recalcular el promedio y la desviación estándar antes de recalcular el estadístico g.

El estadístico se calcula:

$$g_{min} = \frac{\bar{x} - x_{min}}{s} \qquad g_{max} = \frac{x_{max} - \bar{x}}{s}$$

Dónde

$\bar{x} =$ Promedio
s= Desviación Estándar
$x_{min} =$ Valor mínimo
$x_{max} =$ Valor máximo

Valores críticos para la prueba de Grubbs

n	gcrit α=0.05	gcrit α=0.01	n	gcrit α=0.05	gcrit α=0.01	n	gcrit α=0.05	gcrit α=0.01
3	1.1531	1.1546	15	2.409	2.7049	80	3.1319	3.5208
4	1.4625	1.4925	16	2.4433	2.747	90	3.1733	3.5632
5	1.6714	1.7489	17	2.4748	2.7854	100	3.2095	3.6002
6	1.8221	1.9442	18	2.504	2.8208	120	3.2706	3.6619
7	1.9381	2.0973	19	2.5312	2.8535	140	3.3208	3.7121
8	2.0317	2.2208	20	2.5566	2.8838	160	3.3633	3.7542
9	2.1096	2.3231	25	2.6629	3.0086	180	3.4001	3.7904
10	2.1761	2.4097	30	2.7451	3.1029	200	3.4324	3.822
11	2.2339	2.4843	40	2.8675	3.2395	300	3.5525	3.9385
12	2.285	2.5494	50	2.957	3.3366	400	3.6339	4.0166
13	2.3305	2.607	60	3.0269	3.4111	500	3.6952	4.0749
14	2.3717	2.6585	70	3.0839	3.471	600	3.7442	4.1214

Tabla 5. Valores críticos del estadístico g.

Si retomamos los datos del primer ejemplo y queremos probar si el valor 945 y 932.8 son atípicos procederíamos de la siguiente manera:

El promedio de todos los datos es 326.4 y la desviación estándar es 3.233. El valor de g max para 945 es 5.744. Para alfa de 0.05 y 200 puntos el valor crítico de tabla es 3.4324 por lo que el valor es un dato atípico.

Eliminando los tres valores de 945 el nuevo promedio es 926.15 y la nueva desviación estándar es 2.28. El valor de g max ahora es 2.92 para el dato de 932.8. Como el valor crítico sigue siendo 3.4324 el valor no puede considerarse atípico.

Si se prueba si el valor mínimo (919.7) es atípico, el valor de g min sería de 2.83 (después de eliminar el valor de 945 que resultó atípico) y mucho menor que el valor crítico de 3.4324, por lo que no es atípico.

Conclusión

Se ha presentado dos métodos gráficos (el histograma y el Diagrama de Caja) para estimar valores atípicos y se ha analizado sus limitaciones. Los métodos de la Desviación Estándar, la Extensión Camp Meidell, la prueba de Nalimov y la de Grubbs se pueden usar con muestras de casi cualquier tamaño. El método desarrollado por Dean y Dixon se puede utilizar con muestras pequeñas y grandes que no pasen de 100 valores, y la prueba de valor Discordante con muestras menores que 50.