CONTACTO:

Eduardo Tenorio G

- edotenorio@yahoo.com
- 301 6118357
- eduardo.tenorio.galeano@gmail.com
- skype: edotenorio

RECHAZO DE DATOS DE UNA SERIE

Recordemos...

- Cuando se escoge el nivel de significación 0,05 (ó 5%), tenemos un 95% de ***confianza*** de que hemos adoptado la decisión correcta y una probabilidad 0,05 de ser falsa.

...

¿Datos átipicos?

Valores son a menudo referidos como outliers, anomalías, **observaciones discordantes**, excepciones, fallas, defectos, aberraciones, **ruido,** errores, daños, sorpresas, novedades, peculiaridades, **contaminantes**, valores atípicos o valores extremos en diferentes dominios de aplicación [Chandola et. al., 2007].

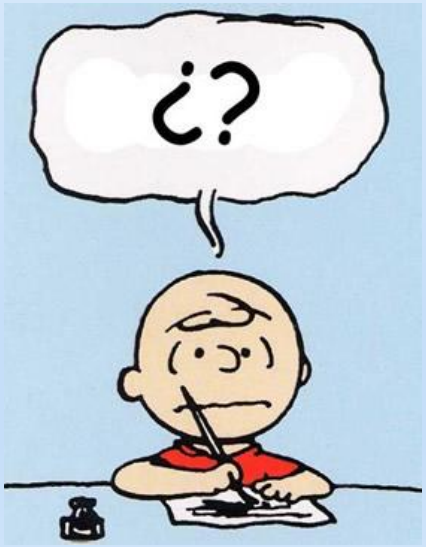
- **Un valor atípico es una observación con un valor que no parece corresponderse con el resto de los valores en el grupo de datos.** *[Mejía Z; Gloria María]*

¿Cómo determina si un valor es realmente un valor atípico y cómo decide si debe continuar o no con el análisis de datos?

- Uno de los problemas en el análisis de datos es **manejar** los valores atípicos dentro de un grupo de datos.

Por lo general surgen dos preguntas:

1) ¿Es este valor **realmente** un valor atípico?
2) ¿**Puedo eliminar** este valor y continuar con el análisis de datos?



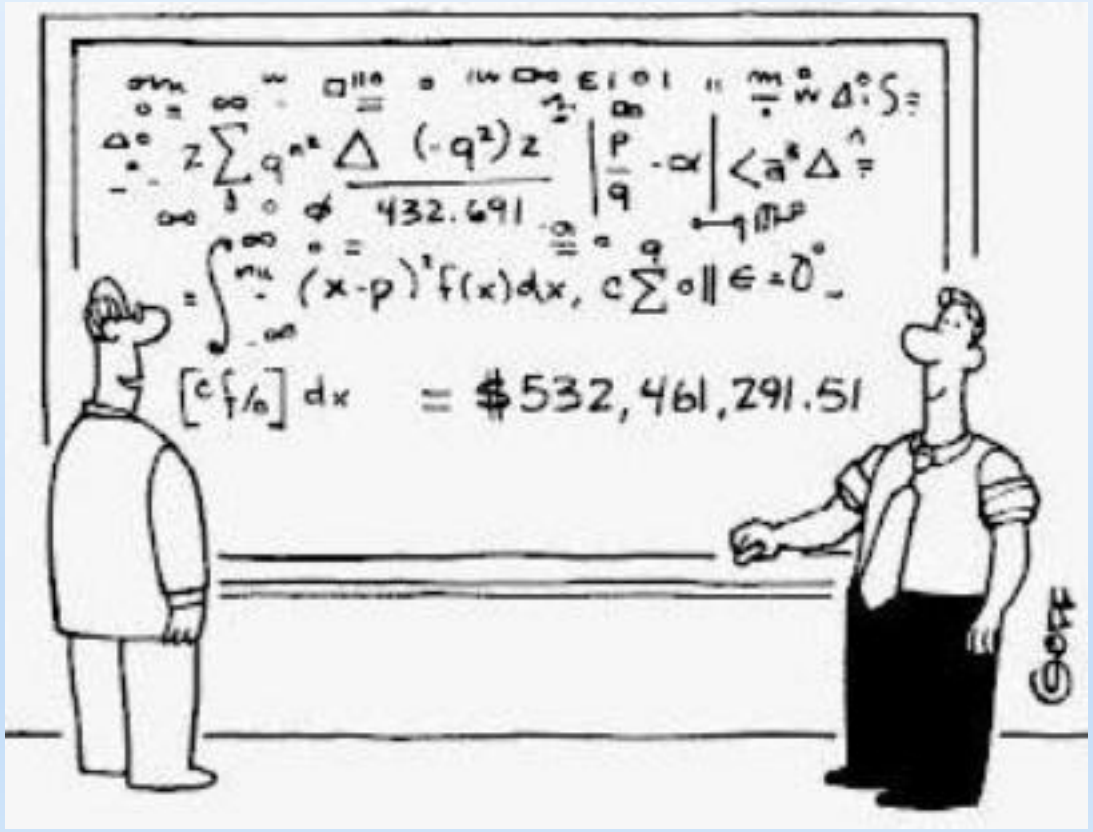
Con respecto a la pregunta 2:

- Debe saberse que las pruebas estadísticas se utilizan para **identificar** valores atípicos, **no para retirarlos** del grupo de datos.
- Técnicamente, una observación **no debe retirarse** a menos que una **investigación** halle una causa probable para **justificar** esta acción

Si en la investigación **no se** encuentra una **causa** probable, **¿qué debe hacerse?**

- Un enfoque sería realizar un análisis de datos **con** el valor atípico y **sin** él. Si las conclusiones son diferentes, entonces si se considera que el valor atípico tiene influencia y esto **debería** indicarse en el **informe**.
- Otra opción es utilizar estimadores rigurosos para caracterizar los grupos de datos, tal como la **mediana** de la muestra en lugar de la **media.**

Pruebas



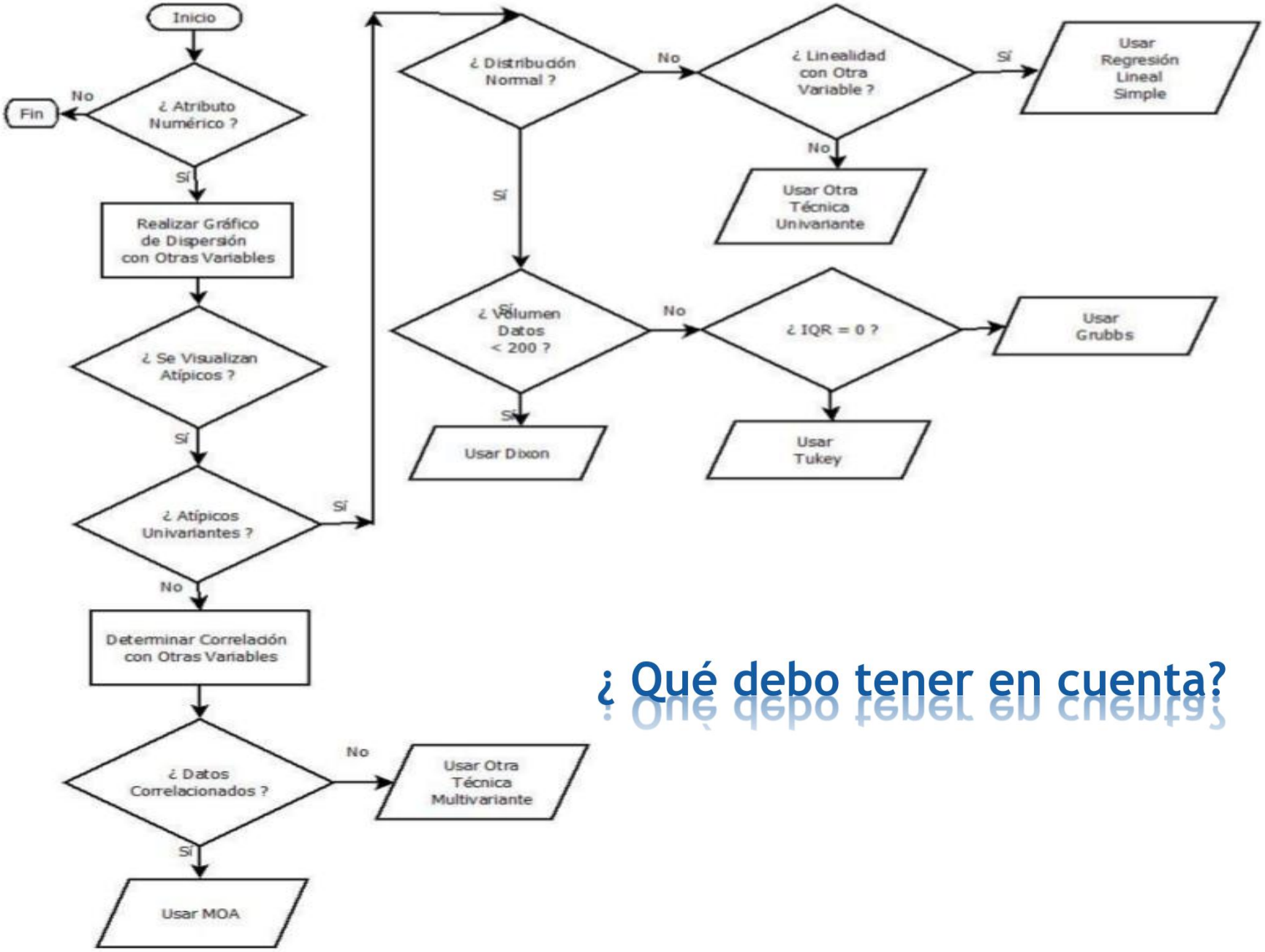
* Prueba de Grubbs.
* Prueba de Dixon.
* Prueba de Tukey.
* MOA.
* Regresión Lineal Simple.
* Criterio de Chauvenet
* Criterio de Peirce.

- Las pruebas estadísticas utilizadas con mayor frecuencia en un grupo de datos son la prueba de **Dixon**, **Tukey** y la prueba de **Grubbs**.

- Algunas de estas técnicas están diseñadas para detectar un **único valor** atípico en un grupo de datos, y por lo tanto **no son** adecuadas para la detección de múltiples valores atípicos.

- Una técnica rigurosa y amplia para identificar eficazmente **múltiples valores atípicos** es el procedimiento para muchos valores atípicos con generalización extrema de la desviación de Student.

¿ Qué debo tener en cuenta?



¿ Qué debo tener en cuenta?

DIXON

La prueba de Dixon utiliza **relaciones** de las diferencias entre datos que parecen atípicos comparados con los valores del grupo de datos.

DIXON

Número de datos	Relación a calcular
n = 3 a 7	r_{10}
n = 8 a 10	r_{11}
n =11 a 13	r_{21}
n = 14 a 24	r_{22}

Fuente: [Taylor y Cihon, 2004]

R	Si x_n es sospechoso	Si x_1 es sospechoso
r_{10}	$\frac{(x_n - x_{n-1})}{(x_n - x_1)}$	$\frac{(x_2 - x_1)}{(x_n - x_1)}$
r_{11}	$\frac{(x_n - x_{n-1})}{(x_n - x_2)}$	$\frac{(x_2 - x_1)}{(x_{n-1} - x_1)}$
r_{21}	$\frac{(x_n - x_{n-2})}{(x_n - x_2)}$	$\frac{(x_3 - x_1)}{(x_{n-1} - x_1)}$
r_{22}	$\frac{(x_n - x_{n-2})}{(x_n - x_3)}$	$\frac{(x_3 - x_1)}{(x_{n-2} - x_1)}$

Fuente: [Taylor y Cihon, 2004]

DIXON

La prueba de Dixon se usa en un número pequeño de observaciones (menor a 26) y detecta elementos que se encuentren sesgados o que son extremos.

Para aplicar la prueba de Dixon se recomienda de un número de observaciones igual o mayor a 8. En este caso se usa r11
que las observaciones sean menores a 10 se utiliza como valor esperado el valor de relación r10

DIXON

Por ejemplo, tomemos los datos 5.3, 3.1, 4.9, 3.9, 7.8, 4.7 y 4.3

Ordenando los datos:

3.1, 3.9, 4.3, 4.7, 4.9, 5.3, 7.8

El tamaño de la muestra es 7, y la relación utilizada es el espacio entre el valor atípico (**7.8)** y su vecino más próximo (**5.3)** dividido por el espacio entre los valores más grandes y más pequeños en el grupo.

Por lo tanto, el índice de Dixon es:

(7.8 – 5.3)/(7.8 – 3.1) = 2.5/4.7 = 0.532

- Este valor se **compara** con un valor crítico de una tabla, y el valor se declara valor atípico si supera ese valor crítico.

$$\textbf{Si } D_{calculado} > D_{tabulado} \quad \text{se rechaza el dato}$$

- El valor tabulado depende del tamaño de la muestra, n, y de un nivel de confianza elegido, que es el riesgo de rechazar una observación válida. La tabla por lo general utiliza niveles de baja confianza tal como 1% o 5%.

- Para un n = 7 y un riesgo del **5%,** el valor en la tabla es **0.507.** El índice de Dixon 0.532 excede este valor crítico, indicando que el valor 7.8 es un valor atípico.

Ejemplo:

Al efectuar una serie de réplicas para determinar el valor obtenido por cinco tácnicos en una muestra se obtuvieron los siguientes resultados. Determinar si la medida **6.0** es un valor rechazable con nivel de confianza del **95%.**

Medida	Valor
1	5.0
2	5.2
3	5.5
4	5.6
5	6.0

1. Se ordenan los datos en orden de valor decreciente
6.0, 5.6, 5.5, 5.2, 5.0

2. Se calcula Q
Q= (6.0-5.6)/ (6.0-5.0) =0.40

3. Se compara Q calculado con Q tabulado para 5 medidas y un nivel de confianza del 90. $Q_{tab}=0.71$
0.40<0.71, luego el valor 6.0 no es rechazable

Dixon Q-test:

N	Q_{crit} (CL:90%)	Q_{crit} (CL:95%)	Q_{crit} (CL:99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

Dixon Q-test:

Sample Size	Low Outlier Test Statistic	High Outlier Test Statistic	Polynomial for Critical value, R_c
3 to 7	$R = \frac{X_2 - X_1}{X_n - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_1}$	$R_c = 1.975 - 0.4994n + 0.5895n^2 + 0.0025n^3$
8 to 10	$R = \frac{X_2 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-1}}{X_n - X_2}$	$R_c = 1.23 - 0.125n + 0.005n^2$
11 to 13	$R = \frac{X_3 - X_1}{X_{n-1} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_2}$	$R_c = 0.90 - 0.03n$
14 to 25	$R = \frac{X_3 - X_1}{X_{n-2} - X_1}$	$R = \frac{X_n - X_{n-2}}{X_n - X_3}$	$R_c = 0.9975 - 0.04268n + 0.000764n^2$
26 to 200	$R = \frac{X_n - \overline{X}}{S_X}$	$R = \frac{\overline{X} - X_1}{S_X}$	$R_c = 2.2795 + 0.025012n - 0.00018427n^2 + 4.61106 \times 10^{-7} n^3$

Fuente: [Davis y McCueen, 2005]

Con ejemplos vamos a R.....

Ejemplo 1:

Dixon test for outliers

data: datos_1
Q = 0.53191, p-value = 0.07633
alternative hypothesis: highest value 7.8 is an outlier

Ejemplo 2:

Dixon test for outliers

data: ejemplo_2
Q = 0.4, p-value = 0.5258
alternative hypothesis: highest value 6 is an outlier

vamos a R.....

Usage

```
dixon.test(x, type = 0, opposite = FALSE, two.sided = TRUE)
```

Arguments

x	a numeric vector for data values.
opposite	a logical indicating whether you want to check not the value with largest difference from the mean, but opposite (lowest, if most suspicious is highest etc.)
type	an integer specyfying the variant of test to be performed. Possible values are compliant with these given by Dixon (1950): 10, 11, 12, 20, 21. If this value is set to zero, a variant of the test is chosen according to sample size (10 for 3-7, 11 for 8-10, 21 for 11-13, 22 for 14 and more). The lowest or highest value is selected automatically, and can be reversed used `opposite` parameter.
two.sided	treat test as two-sided (default).

Details

The p-value is calculating by interpolation using `qdixon` and `qtable`. According to Dixon (1951) conclusions, the critical values can be obtained numerically only for n=3. Other critical values are obtained by simulations, taken from original Dixon's paper, and regarding corrections given by Rorabacher (1991).

GRUBBS

La prueba de Grubbs utiliza una estadística de prueba, T, que es la diferencia absoluta entre el valor atípico, X_O, y el promedio de la muestra (X) dividida por la desviación estándar de la muestra, s.

Para el ejemplo anterior, el promedio de la muestra es = 4.86 y la desviación estándar de la muestra es = 1.48. La estadística calculada de la prueba es:

$$T = |X_o - \bar{X}|/s = |7.8 - 4.86|/1.48 = 1.99$$

GRUBBS

Para un n = 7 y un riesgo del 5%, el valor tabulado es 1.938 y el $T_{Calculado}$ = 1.99 excede este valor crítico, indicando que el valor 7.8 es un valor atípico.

TEST DE GRUBB PARA DATOS SOSPECHOSOS

Recomendado por las normas ISO

$$G = \frac{\text{Valor Sospechoso} - X}{S}$$

(Con el valor sospechoso incluido)

Si Gcalculada > Gtabulada el valor sospechoso se rechaza

TEST Q DE DATOS SOSPECHOSOS

Aceptar o rechazar un resultado anómalo (outlier)

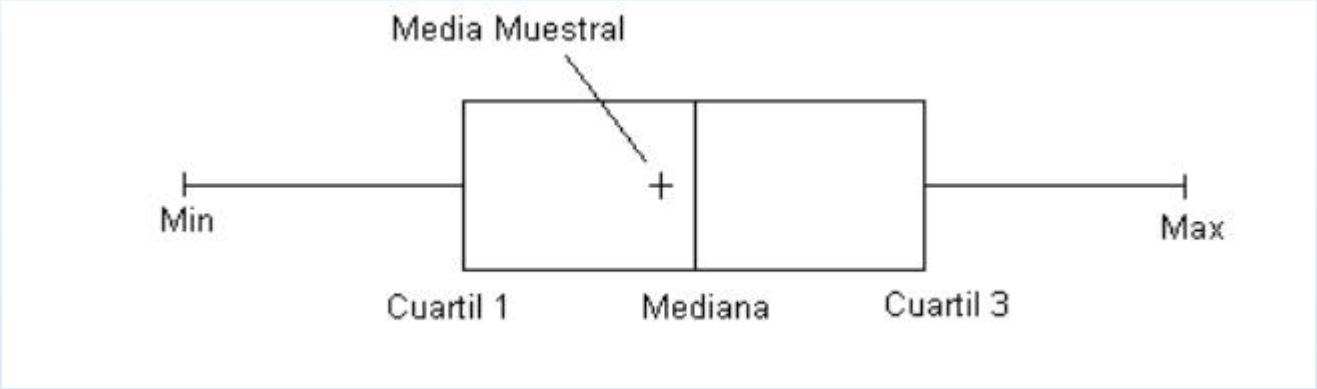
Normalmente se producen al cometer errores o fallos en la metodología aplicada.

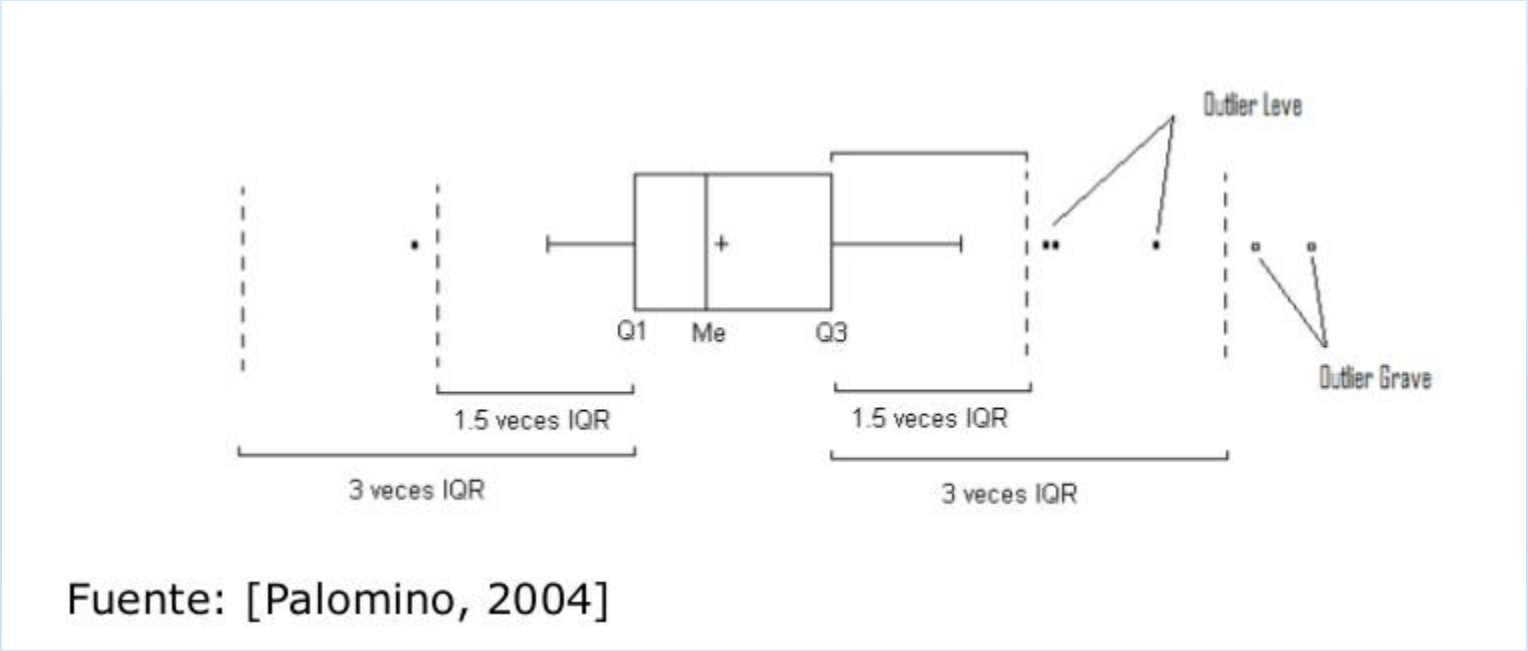
Se ordenan los datos en forma creciente y se calcula Q

$$Q = \frac{\text{desvío}}{\text{recorrido}} = \frac{\text{Diferencia entre el dato sospechosos y su vecino más cercano}}{\text{Diferencia numérica entre el dato de mayor valor y el de menor valor}}$$

Si Qcalculada > Qtabulada el dato se rechaza

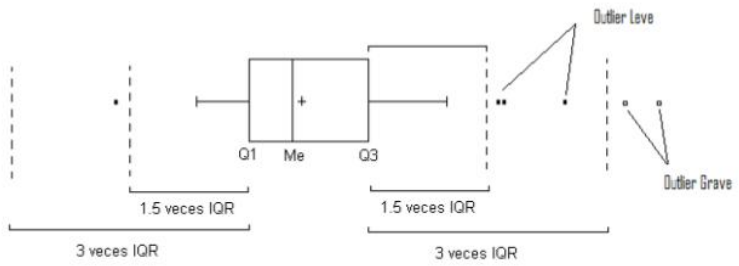
Prueba de Tukey..





Fuente: [Palomino, 2004]

Prueba de Tukey..



Fuente: [Palomino, 2004]

Para la detección de los valores atípicos, la longitud máxima de cada uno de los bigotes es de K = 1,5 veces el rango intercuartil (IQR) es decir 1.5x(Q3-Q1) por encima y por debajo de los cuartiles. Las observaciones fuera de los bigotes son dibujadas separadamente y etiquetadas como valores atípicos. El método de Tukey utiliza un K=3 adicionalmente del K =1.5, las observaciones que están entre 1.5 y 3 veces el rango intercuartil reciben el nombre de atípicos leves. Las observaciones que están más allá de 3 veces el rango intercuartil se conocen como valores atípicos extremos. En la figura 14 se muestra un diagrama de cajas y bigotes con valores atípicos leves y graves.

CONCLUSIONES

- La ASTM E178, Práctica para manejar observaciones de valores atípicos, contiene muchos procedimientos estadísticos para realizar pruebas de valores atípicos. En esta norma se proveen otros criterios para valores atípicos únicos, así como pruebas para valores atípicos múltiples, y la norma también da pautas para la elección de la prueba.
- Una referencia más amplia para la prueba de valores atípicos es el libro Valores atípicos en datos estadísticos, publicado por Wiley. Otra referencia útil y más práctica es el Volumen 16 de la Sociedad Estadounidense de Calidad (ASQ) "Referencias básicas para el control de calidad, técnicas estadísticas"
- Cómo detectar y manejar valores atípicos", ASQC Quality Press.
- En la práctica E178 de ASTM se indican otras referencias.

AINSER
ASOCIACIÓN DE INGENIEROS
AL SERVICIO

¿preguntas?

Ing. Eduardo Tenorio G

duardo Tenorio G:

- Eduardo Tenorio G
- edotenorio@yahoo.com
- 301 6118357
- eduardo.tenorio.galeano@gmail.com
- skype: edotenorio

Gracias

REFERENCIAS

- Eurolab España. P.P. Morillas y colaboradores. Guía Eurachem: La adecuación al uso de los métodos analíticos – Una Guía de laboratorio para la validación de métodos y temas relacionados (1a ed. 2016).
- NTC-3529-1,2,3,4,5:1998 Exactitud (veracidad y precisión) de los métodos de medición y de los resultados.
- IISO-13528:2015 Generalidades: Métodos estadísticos para utilizar en programas de ensayos de aptitud mediante comparación interlaboratorios