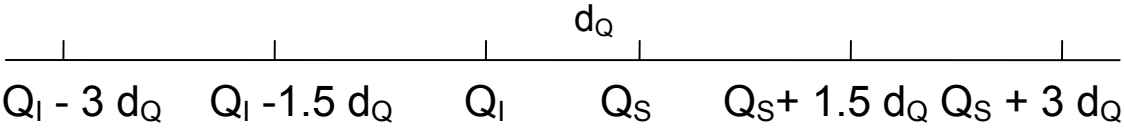
Identificación de valores atípicos.

Utilizamos una medida de dispersión que sea insensible a los valores atípicos, la distancia intercuartil y definimos puntos de corte para detectar outliers:

$$\text{Valla Interna Inferior} = Q_I - 1.5\, d_Q$$
$$\text{Valla Interna Superior} = Q_S + 1.5\, d_Q$$
$$\text{Valla Externa Inferior} = Q_I - 3\, d_Q$$
$$\text{Valla Externa Superior} = Q_S + 3\, d_Q$$



$$\text{VALOR ADYACENTE INFERIOR (VAI)} = \begin{cases} \text{valor más cercano, mayor o igual,} \\ \text{a la valla interna inferior} \end{cases}.$$

$$\text{VALOR ADYACENTE SUPERIOR (VAS)} = \begin{cases} \text{valor más cercano, menor o igual,} \\ \text{a la valla interna superior.} \end{cases}$$

Si no hay valores atípicos: VAI = mínimo VAI = máximo

BOXPLOTS.

El boxplot es la representación gráfica de la mediana, los cuartiles, los valores adyacentes y los valores externos moderados o severos.

Permite extraer los siguientes aspectos del lote:

Posición del centro - Dispersión - Asimetría - Longitud de la cola
Puntos que yacen fuera del conjunto.

n = 15, (n+1)/ 4 = 16 /4 = 4

El boxplot se construye dibujando:

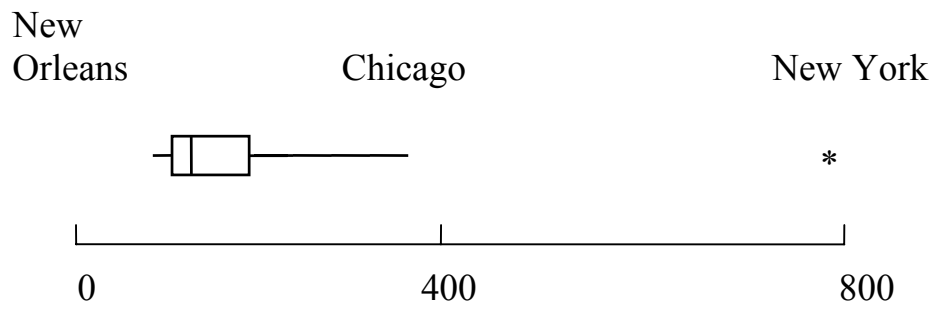
i) una caja cuyos extremos son los cuartiles (Q_I =74) y (Q_S=200) y con una barra vertical en la mediana (88),

ii) una línea de cada extremo de la caja hasta el corresp. valor adyacente (VAI = 63 VAS = 355),

iii) los valores que caen fuera de las vallas internas pero dentro de las externas son outliers moderados (no hay),

iv) los valores que caen fuera de las vallas externas son outliers severos (New York).

Ciudad	Población(10000)
New York	778
Chicago	355
Los Angeles	248
Philadelphia	200
Detroit	167
Houston	94
Baltimore	94
Cleveland	88
Washington, DC	76
St. Louis	75
San Francisco	74
Milwakee	74
Boston	70
Dallas	68
New Orleans	63



Boxplot de la población de las 15 ciudades más grandes de USA.

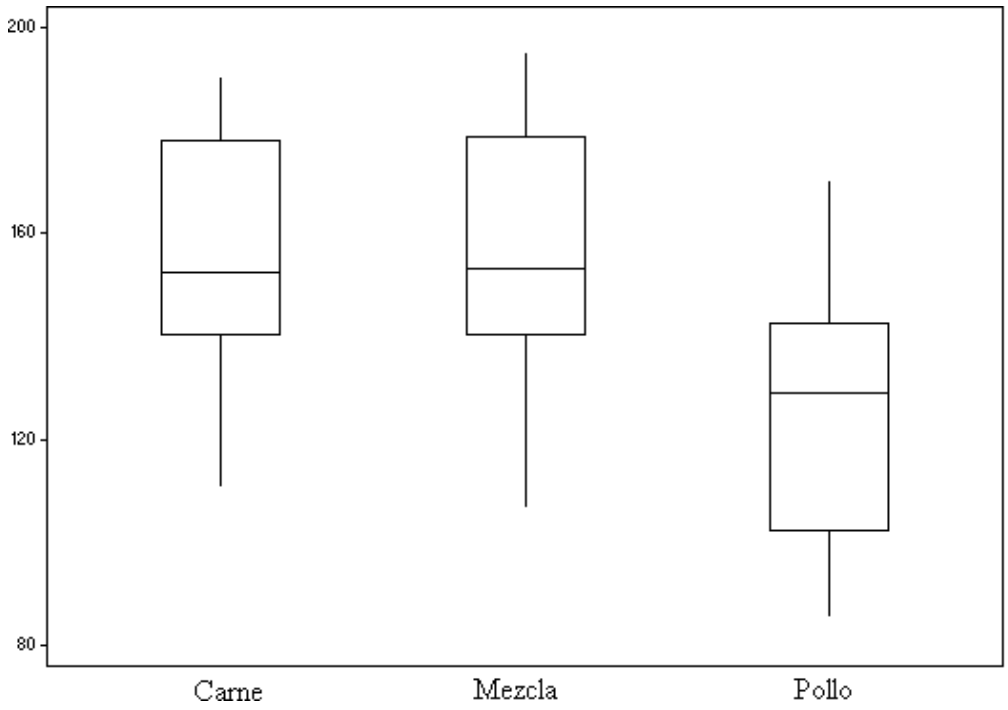
Del boxplot para la población de las 15 ciudades más grandes de USA vemos que el lote es pesadamente asimétrico y hay un punto atípico.

Resistencia del Boxplot

Un gráfico similar podría construirse en base a la media y el desvío muestrales. Tal gráfico carecería de resistencia.

Comparación de lotes

Boxplots del contenido calórico de tres tipos de salchichas



Diagramas-tallo hoja de los datos de calorías en diferentes clases de salchichas.

Vacuno		Mezcla		Pollo	
Tallo	Hojas	Tallo	Hojas	Tallo	Hojas
8		8		8	67
9		9		9	49
10		10	7	10	226
11	1	11		11	3
12		12		12	9
13	1259	13	5689	13	25
14	1899	14	067	14	2346
15	2378	15	33	15	2
16		16		16	
17	56	17	2359	17	0
18	146	18	2		
19	00	19	015		

De los Box-Plots:
Las salchichas de pollo, como grupo, contienen menos calorías que las de carne o las de mezcla: la mediana del contenido calórico de las de pollo está por debajo del cuartil inferior de las otras distribuciones. Todos los tipos muestran una gran dispersión entre marcas; las salchichas de pollo no garantizan una comida de bajas calorías.

De los diagramas Tallo-Hoja:
Para los datos de "mezcla" vemos que se distinguen claramente dos grupos de marcas, la distribución tiene dos picos y un outlier en la cola inferior.

Los cuartiles, Ci=139.50 y Cs=179.75, están aproximadamente en el centro de cada uno de los grupos, de manera que gran parte de la distancia intercuartil (d_c) está dada por la distancia entre los grupos. Por esta razón el 1.5* d_c que se utiliza para graficar el box-plot no distinguió al outlier.

Aunque en el diagrama correspondiente a las salchichas de pollo no se observan dos grupos separados, como en "vacuno" y "mezcla", pueden verse claramente dos picos.

Esta *distribución bimodal* también sugiere la presencia de dos grupos en los datos.

Ni la media ni la mediana dan una buena información sobre este tipo de datos porque no está presente en ellos un claro centro.

Valores atípicos:

Ejemplo 1:
En 1985 los científicos británicos anunciaron un agujero en la capa de ozono de la atmósfera terrestre sobre el polo sur.

El reporte de los británicos fue descartado al comienzo pues estaba basado en instrumentos terrestres enfocados hacia arriba. Observaciones más completas, obtenidas por instrumentos satelitales mirando hacia abajo, no habían mostrado nada inusual.
Luego, un análisis más exhaustivo de las mediciones satelitales, reveló que las lecturas de ozono en el polo sur eran tan bajas que el programa de computadora que las analizaba las había suprimido automáticamente como outliers en forma equivocada.

Se reanalizaron las lecturas desde 1979. Éstas mostraron un agujero de tamaño creciente en la capa de ozono que no tenía explicación.

Ejemplo 2:Mediciones obtenidas por Newcomb entre Julio y Septiembre de 1882.

28	22	36	26	26	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

¿qué variable ha sido medida?

- Newcomb midió cuánto tardó la luz en llegar, desde su laboratorio sobre el río Potomac a la base del monumento a Washington y volver, una distancia total de 7400 metros.
- es necesario tener la descripción del instrumento
- juzgar si la variable medida es la adecuada (conocimiento experto)
- sobre el campo particular en estudio.

Por ejemplo Newcomb construyó aparatos nuevos y complicados para medir el tiempo en que pasaba la luz. Nosotros aceptamos el juicio de los físicos sobre que este instrumento es adecuado para su propósito y más preciso que instrumentos anteriores.

Codificación: La primera medición del tiempo de paso de la luz era 0.000024828 segundos. Corremos al punto decimal nueve lugares a la derecha, obteniendo 24828 y luego registramos únicamente **el desvío respecto de 24800**. Luego 28 es la versión corta de 0.000024828 y -2 se corresponde con 0.000024798.

Variación

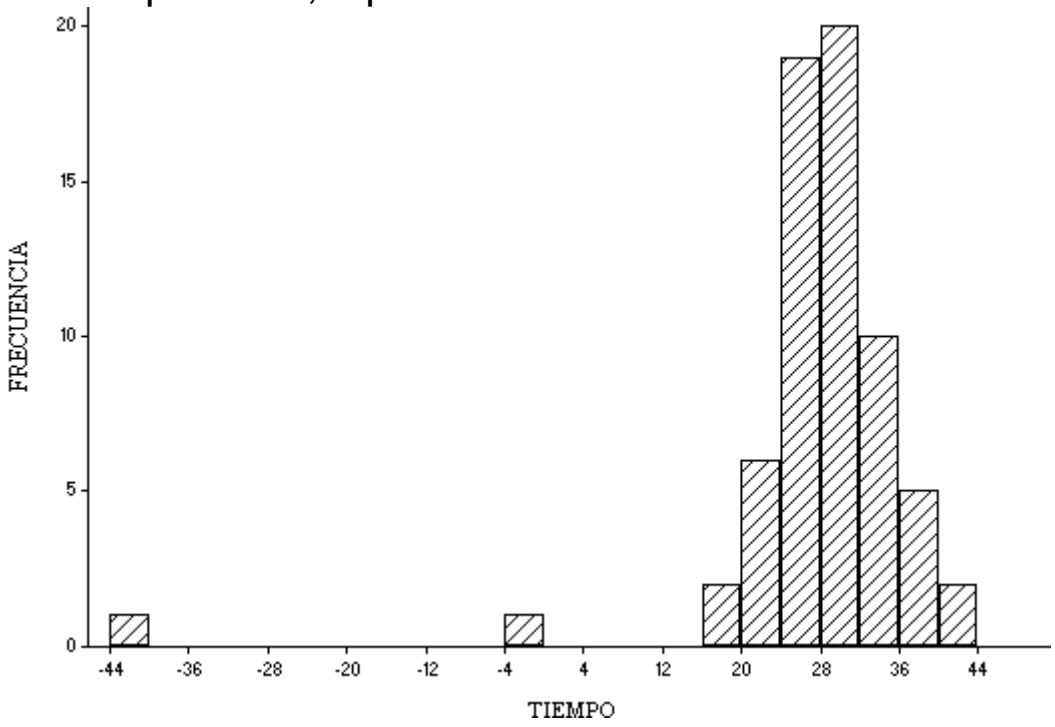
Los aparatos cambian levemente con la temperatura, la densidad de la atmósfera cambia día a día y así siguiendo.

Incluso los mejores experimentos producen resultados variables.

Esta es la razón porque Newcomb tomó muchas mediciones en vez de una.

> En general, el promedio de varias observaciones es menos variable que el de una única observación.

Poniéndonos en lugar de Newcomb, estamos tentados de calcular el promedio de los tiempos de pasaje de la luz, convertir este tiempo en una estimación nueva y mejor de la velocidad de la luz y correr, para hacernos una reputación, a publicar el resultado. !!PELIGRO!!



Histograma de las 66 mediciones de Simon Newcomb

Un dato atípico en la brillantez vista por un satélite de vigilancia puede representar el lanzamiento de un misil.

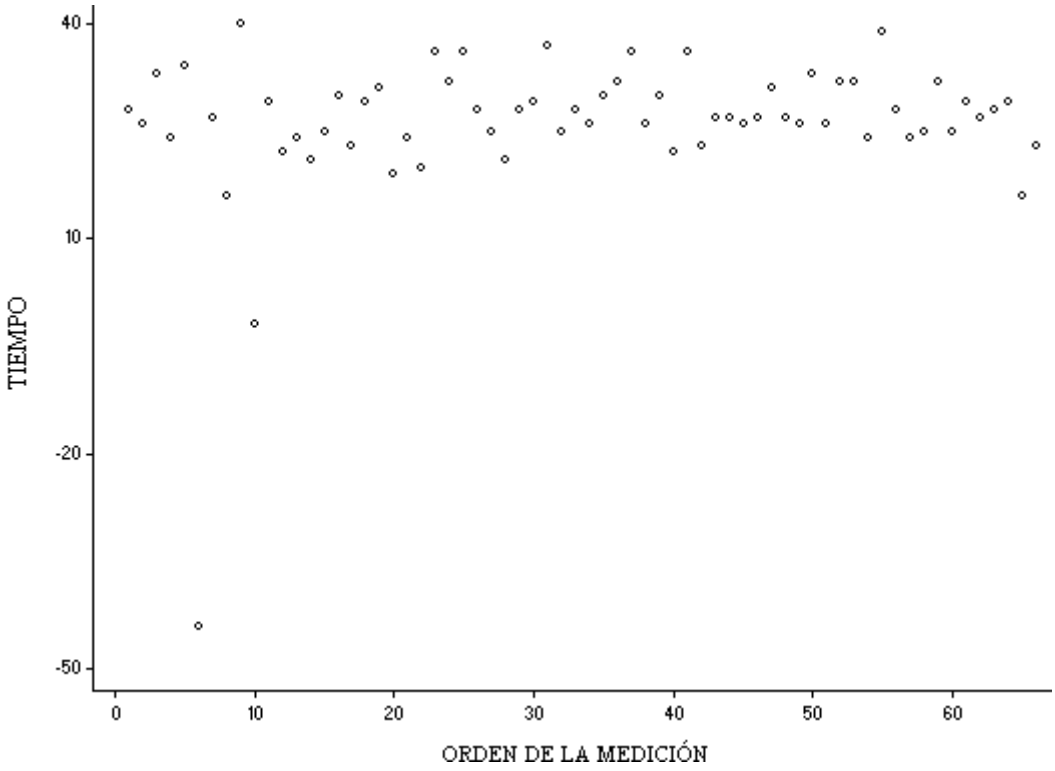
Un dato atípico de las mediciones de actividad eléctrica en un detector utilizado en física de altas energías puede ser evidencia de una nueva partícula elemental.

En tales casos la distribución general simplemente provee un patrón de referencia sobre el cual sobresalen los eventos extraordinarios.

Cuando los datos atípicos son inesperados e indeseados se debería hallar una causa clara para cada outlier, como la falla del equipo durante el experimento o un error en la transcripción de los datos, en esos casos, se puede corregir o eliminar el dato.

Cuando no se encuentra ninguna causa es muy difícil tomar una decisión.

Newcomb finalmente eliminó el peor outlier (-44) pero retuvo el otro. La media de todas las 66 observaciones es 26.21; la media de las 65 observaciones retenidas es 27.29. El gran efecto del único valor -44 sobre la media es la razón para eliminarlo.



Este gráfico sugiere levemente que la variabilidad (dispersión vertical) es decreciente con el tiempo. Quizás, a medida que ganó experiencia, Newcomb se volvió más experto en el uso de su equipo.

Los efectos de aprendizaje como el que muestran los datos de Newcomb son muy frecuentes y deben ser tenidos en consideración.
Si dejamos las primeras 20 observaciones de Newcomb para el aprendizaje, la media de las 46 restantes resulta 28.15. Las mejores

mediciones modernas sugieren que el “verdadero valor” para el tiempo de paso de la luz del experimento de Newcomb es 33.02.
Eliminar los outliers ó fijar un período de aprendizaje, acercan los resultados al verdadero valor. Pero si es posible, siempre, hay que hallar la razón de un outlier.

RESUMEN

- Una **medida resistente** no se ve afectada por cambios en los valores numéricos de una pequeña proporción de la cantidad total de observaciones, sin importar cuánto cambien estos valores.
- El **centro** de una distribución es medido por la **media**, la **media** α **podada** ó la **mediana**. La media es el promedio aritmético de todos los datos. La media α podada es el promedio aritmético de los datos excluidos el 100*α% de los valores mayores y el 100*α% de los valores menores. La mediana es el punto medio de los datos ordenados.
- La distancia intercuartil provee una medida resistente de la **dispersión** o **variabilidad** de la distribución. Los **cinco números resumen**, dados por la mediana, los cuartiles, el máximo y el mínimo proveen una descripción rápida de la forma global de una distribución.
- Los **Boxplots**, basados en los cinco números resumen, son útiles para comparar varias distribuciones. Las **vallas internas** y **externas** son útiles para identificar potenciales valores atípicos (outliers).
- La **varianza muestral** s^2 y especialmente su raíz cuadrada, el desvío estándar DS, son medidas muy usuales, pero no resistentes, de la dispersión de los datos alrededor de la media.