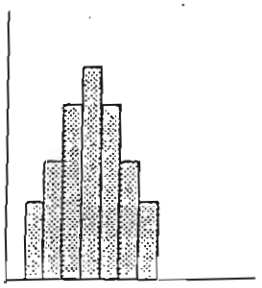
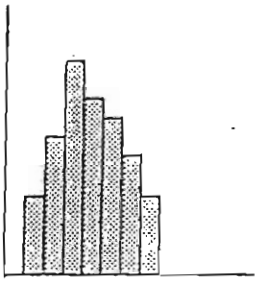
simétrica. Si la distribución de frecuencias tiene a la derecha una cola más larga que a la izquierda se dice que es **asimétrica o sesgada a la derecha**; en caso contrario, se dice **asimétrica o sesgada a la izquierda**.



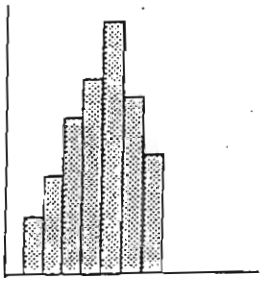
$\bar{x}=Me=Mo$

Simétrica



$Mo<Me<\bar{x}$

Asimétrica a derecha



$\bar{x}<Me<Mo$

Asimétrica a izquierda

1.5.4.1. Sesgo.

El sesgo viene dado por la diferencia entre la media aritmética y la moda y se hace adimensional dividiéndola entre la desviación estándar: $\upsilon = \dfrac{\bar{x} - Mo}{s_n}$. Este estadístico tiene sentido en distribuciones en que la moda es única. Mide el grado de asimetría del siguiente modo: si la distribución es simétrica o insesgada entonces $\upsilon = 0$; si la distribución es asimétrica a la derecha entonces $\upsilon > 0$; si la distribución es asimétrica a la izquierda entonces $\upsilon < 0$.

Ejemplo 1.23: Vamos a calcular el sesgo de la distribución de frecuencias del ejemplo de los pollitos para el caso de los datos sin agrupar. Sabemos que $\bar{x} = 9.8778$ días, Mo=8 días y $s_n^2 = 8.7295$ días^2, con lo que $s_n = \sqrt{8.7295} = 2.9546$ días. Así, $\upsilon = \dfrac{9.8778 - 8}{2.9546} = 0.6356 > 0$. Por lo tanto, la distribución de frecuencias es asimétrica a la derecha. Esta conclusión ya la conocíamos al representar gráficamente el diagrama de barras de esta distribución.

ANEXO 1: DIAGRAMAS DE CAJA

Ya tenemos todos los elementos para poder explicar qué es y cómo se construye un diagrama de caja. El **diagrama de caja** es una representación gráfica de un conjunto de datos que facilita la percepción visual de la posición y extensión de los mismos. También permite identificar los **outliers** o valores raros o extremos. Es especialmente útil cuando se desean comparar dos o más conjuntos de datos. Estos diagramas son construidos por cualquier software especializado en estadística, pero vamos a explicar cómo se realiza su construcción. Los pasos son los siguientes:

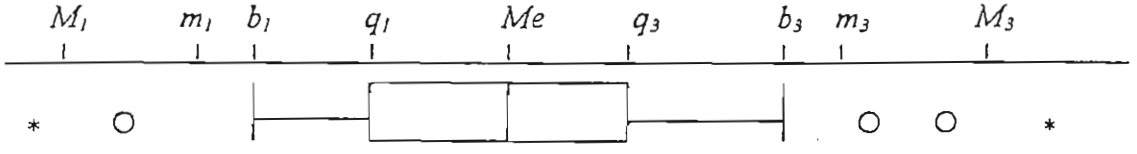
1) Construir una escala de referencia horizontal o vertical.
2) Determinar la mediana muestral, el primer y tercer cuartil y el rango intercuartílico.
3) Determinar los puntos m_1 y m_3 llamados "**vallas interiores**", de la forma:
   $m_1 = q_1 - 1.5 \cdot RI$
   $m_3 = q_3 + 1.5 \cdot RI$

Estos puntos se utilizarán para identificar los outliers.

4) Determinar los puntos b_1 y b_3 denominados "**valores adyacentes**". El punto b_1 es el valor más cercano a m_1 sin que su valor esté por debajo de m_1. El punto b_3 es el valor más cercano a m_3 sin que su valor esté por encima de m_3.
5) Determinar dos puntos M_1 y M_3 llamados "**vallas exteriores**", de la forma:
   $M_1 = q_1 - 2 \cdot (1.5) \cdot RI$

$M_3=q_3+2\cdot(1.5)\cdot RI$

6) Situar los puntos hallados hasta ahora sobre la escala horizontal o vertical.
7) Construir una caja con los extremos en q_1 y q_3 con una línea interior dibujada en la mediana.
8) Indicar los valores adyacentes mediante unas patillas y conectarlos a la caja.
9) Situar los datos puntuales que estén entre las vallas interior y exterior representados mediante círculos abiertos. Estos puntos se consideran como outliers moderados.
10) Indicar los datos puntuales que caen fuera de las vallas exteriores mediante asteriscos. Estos puntos se consideran como outliers extremos.



Ejemplo 1.23: Se ha realizado un estudio sobre la amnesia postraumática tras una lesión en la cabeza. Una variable estudiada es el tiempo de hospitalización en días. Los datos obtenidos son los siguientes para 21 pacientes:

12	27	32	8	35	36	40
20	30	40	40	47	89	108
40	42	45	61	52	41	50

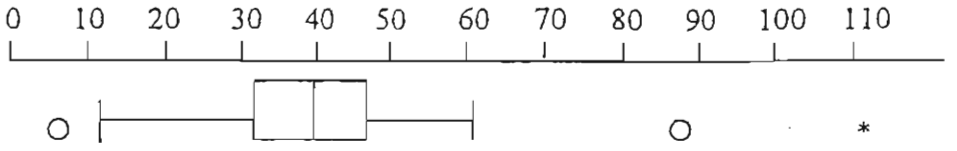
Vamos a construir el diagrama de caja correspondiente. Primero calculamos la mediana y el primer y tercer cuartil:

Nº de días x_i	f_i	F_i
8	1	1
12	1	2
20	1	3
27	1	4
30	1	5
32	1	6
35	1	7
36	1	8
40	4	12
41	1	13
42	1	14
45	1	15
47	1	16
50	1	17
52	1	18
61	1	19
89	1	20
108	1	21
Total	21	

Como n=21 entonces $n/2$=10.5, luego Me=40. Por otra parte $n/4$=21/4=5.25, luego q_1=32. Además, $3n/4$=3.21/4=15.75, luego q_3=47. Con todo esto, el rango intercuartílico es RI=47-32=15. Las vallas interiores son: m_1=32-1.5·(15)=9.5 y

m_3=47+1.5·(15)=69.5. Los valores adyacentes son b_1=12 por ser el valor superior a 9.5 más cercano a él y b_3=61 por ser el valor inferior a 69.5 más cercano. Las vallas exteriores son: M_1=32-2·(1.5)·(15)=-13 y M_3=47+2·(1.5)·15=92.

El conjunto de dos outliers moderados, el 8 y el 89, por estar entre las vallas interiores y exteriores. El punto 108 es un outlier extremo.



La situación de la línea central de la caja es una indicación de la forma de la distribución. Si la línea está mal centrada, sabremos que la distribución está sesgada en la dirección del extremo más largo de la caja.

Puede demostrarse que al muestrear a partir de una distribución normal, sólo aproximadamente 7 de cada 1000 valores caerán fuera de las vallas interiores. Puesto que estos valores son muy inusuales, se consideran outliers. Los outliers deben tratarse con cuidado pues su presencia puede tener un impacto crucial sobre la media aritmética, la varianza y la desviación típica. Cuando se encuentre un outlier, debería considerarse su causa: ¿es una valor mal registrado?, ¿es el resultado de algún error o accidente en la experimentación?, ¿es legítimo?. En los primeros dos casos pueden borrarse estos datos y completar el análisis con los datos restantes. En el último caso, se sugiere que se dé a conocer la presencia del outlier y que se calculen los estadísticos con y sin el outlier. De esta forma el investigador, que es el experto en la materia, puede tomar la decisión de incluir o no el outlier en futuros análisis.