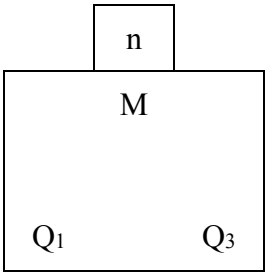
Práctica 5 de Estadística

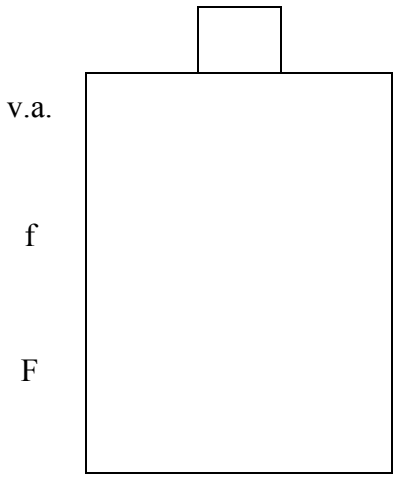
Análisis de una variable medible.
Gráficos caja.

El gráfico caja es una representación gráfica que proporciona información sobre una distribución muestral complementaria a la dada por los histogramas. Este gráfico presta especial interés sobre las colas de la distribución, indicando qué valores pueden considerarse atípicos. Pero también se representan en este gráfico la mediana y los cuartiles, con lo que el gráfico proporciona información tanto de localización como de dispersión. El método está basado en comparar la distribución de frecuencias muestral con aquella que debería corresponder a una distribución normal con recorrido intercuartílico igual al de la muestra. Está basado en el "Principio de Windsor", el cual establece que, en la práctica, todas las distribuciones de frecuencia son normales en el centro.

5.1 Construcción del gráfico caja

La información necesaria para construir el gráfico caja de una muestra puede organizarse de forma conveniente utilizando una caja resumen, que contiene información sobre la mediana y los cuartiles, y añadiendo una segunda caja como se muestra a continuación:

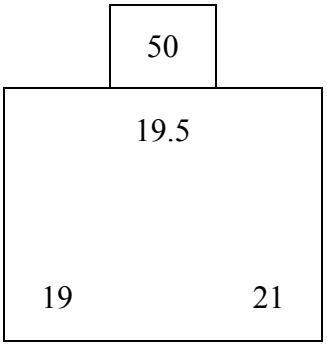




Así, si consideramos la variable edad de los alumnos de una clase de 50:

Edad	17	18	19	20	21	22	23	24	29
Alumnos	2	5	18	9	5	5	3	2	1

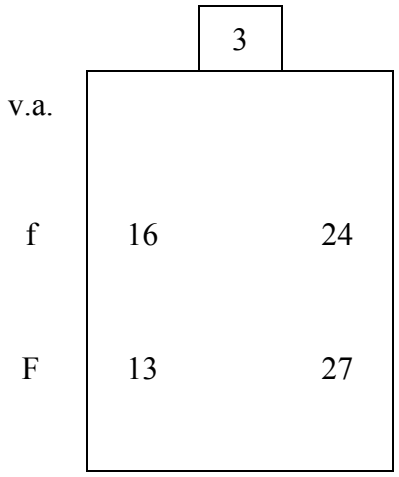
Necesitamos conocer la mediana M = 19.5 y los cuartiles $Q_1 = 19$ y $Q_3 = 21$.
El intervalo intercuartílico es IQR = 2.



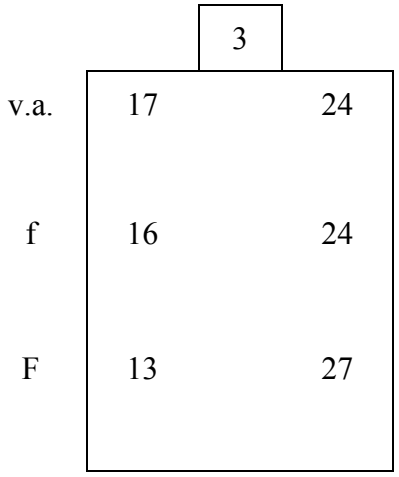
Ahora procederemos a llenar la caja inferior. La caja superior contendrá un factor de escala 1.5 x IQR. Así, para este ejemplo:

$$1.5 \times 2 = 3$$

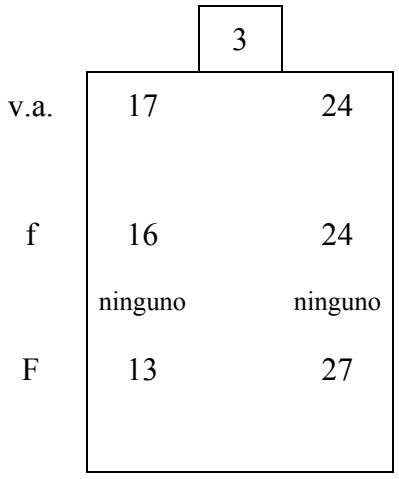
Este valor es restado de $Q_1 = 19$ para obtener la valla inferior, f = 19 - 3 = 16, y de nuevo restamos de este valor (f) para obtener la valla exterior inferior, F = 16 - 3 = 13. Las vallas interior y exterior superiores se obtienen añadiendo el factor de escala a $Q_3 = 21$ obteniéndose f = 24 y F = 27. Estos valores se colocan en el diagrama como sigue:



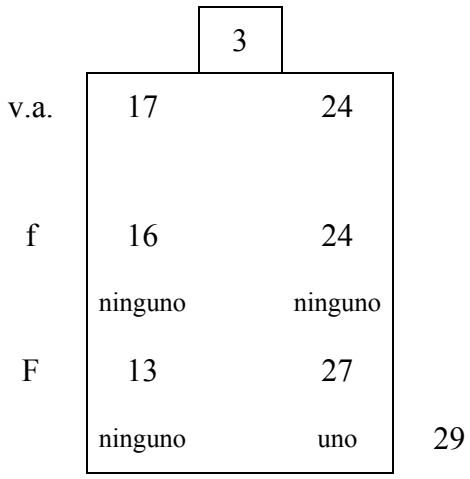
Una vez establecidas estas fronteras, el resto de la información procede de los datos de la variable (es preferible tenerlos ordenados de mayor a menor). Los valores adyacentes (v.a.) son los valores más extremos de la región comprendida entre los cuartiles respectivos y las vallas interiores (incluidas). Así, 17 es el valor mínimo entre 16 y 19, mientras que el valor máximo entre 21 y 24 es 24.



Todas las observaciones pertenecientes a las regiones comprendidas entre la valla interior y exterior serán llamadas atípicas o medianamente atípicas. El hecho de que no haya observaciones entre 13 (incluido) y 16 se indica en el diagrama siguiente. Tampoco hay observaciones mayores que 24 y menores o iguales que 27.



Las observaciones fuera de las vallas exteriores F son llamadas extremas o extremadamente atípicas. No hay datos menores que 13 y tenemos un dato mayor que 27, en concreto 29.



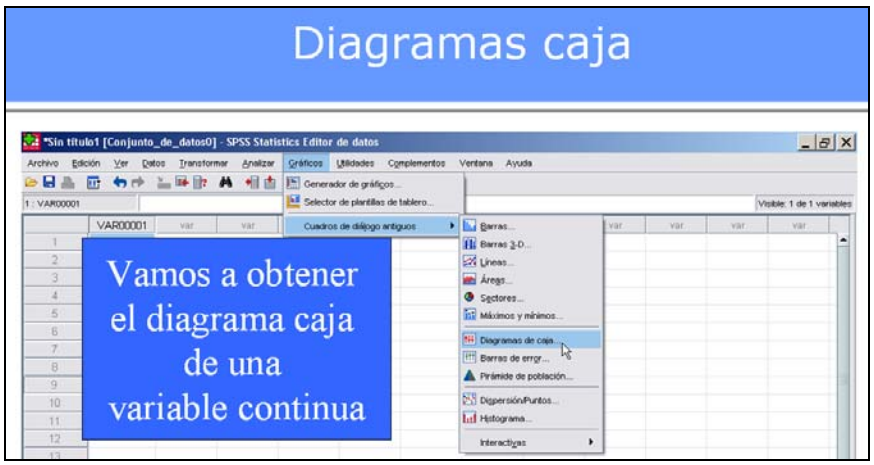
A partir de aquí se construye el gráfico caja. Se requiere para ello una escala que englobe el rango total de los datos. La caja termina en Q_1 y Q_3 con una recta interior trazada en la mediana. A continuación se unen los valores adyacentes a la caja mediante rectas. Finalmente se marcan los valores atípicos y extremos.

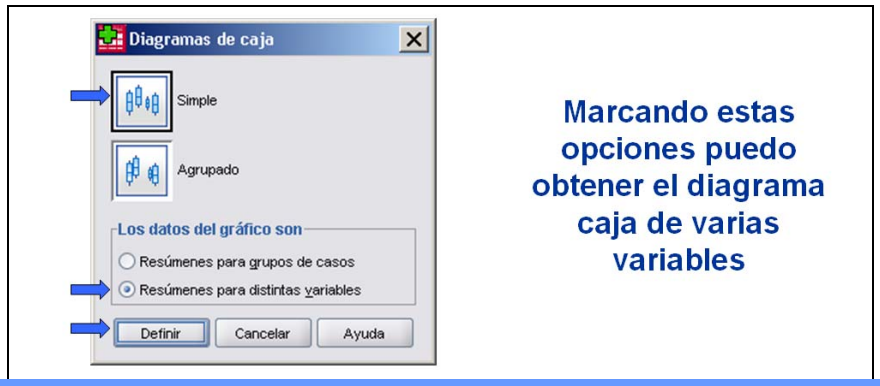
5.2 Interpretación del gráfico caja

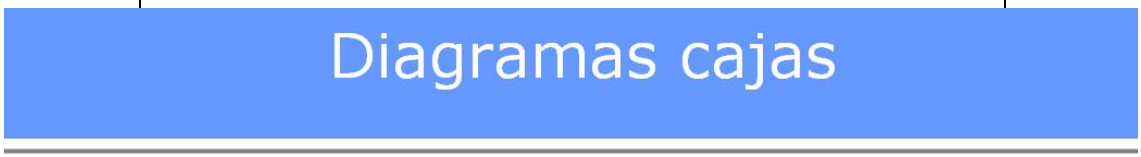
El centro (posición central) de la distribución se representa por la recta mediana. La dispersión central de la distribución viene medida por la longitud de la caja, la cual contiene el 50% de las observaciones, 25% a cada lado de la mediana, la cual informa por tanto de la forma de la parte central. La longitud de los bigotes informan sobre la simetría de las colas; los valores atípicos alargan esas colas.

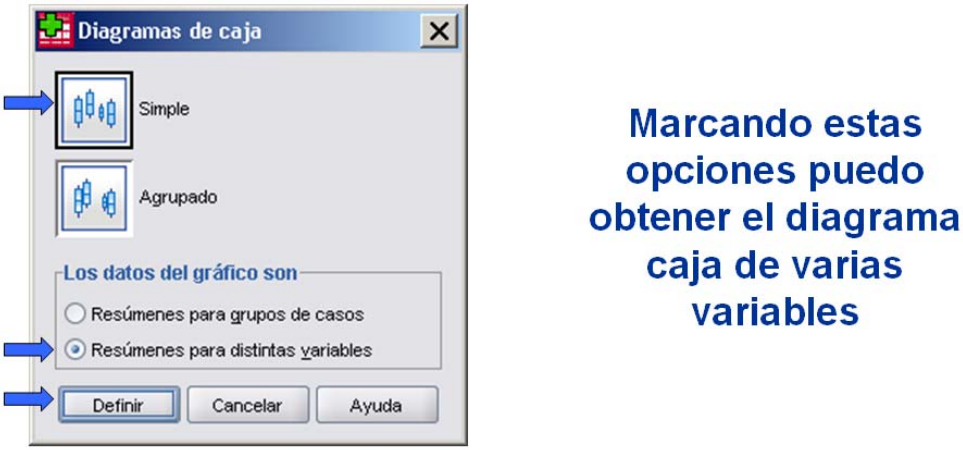
5.3 Manejo de SPSS

Para calcular el gráfico caja escogeremos la opción ‘Gráficos>cuadro de diálogo antiguos>Diagramas de caja’ y seleccionaremos la variable.









5.4 Ejercicios

1. Se desea comparar la calidad del servicio ADSL de tres proveedores distintos. Para ello se dispone de tres ordenadores iguales con la misma configuración sobre los que se ha realizado pruebas en distintos momentos para medir la velocidad de conexión que proporcionan los tres proveedores analizados en Mbps. Los resultados obtenidos son los siguientes:

PROVEEDOR UNO	PROVEEDOR DOS	PROVEEDOR TRES
18,50	15,50	16,70
18,90	14,80	21,30
21,10	15,70	19,20
17,90	17,40	19,70
16,80	16,30	18,55
17,70	17,20	19,50
21,05	16,05	21,10
18,50	19,10	15,35
19,75	15,60	11,30
20,10	15,80	18,90

a) Obtener los gráficos caja para los resultados de los tres proveedores de Internet.
b) ¿Cuál de ellos tiene mayor dispersión? ¿Son simétricos? Indica los valores extremos y atípicos si los hubiera.
c) ¿Qué proveedor es más fiable? ¿Cuál proporciona una mayor velocidad media?

2. Después de la jornada 22 en la liga 2017/2018, los puntos de la clasificación quedan como sigue:

Puntos	46	44	44	41	38	38	34	31	30	30	29	29	29	29	29	28	28	24	24	17

a) ¿Cuál es el tamaño de la muestra? Calcular la media, mediana, desviación típica y cuartiles.
b) Construir el gráfico caja y comentar los resultados (dispersión, valores extremos, simetría…).
c) ¿Cómo interpreta el hecho de que el primer cuartil está muy próximo a la mediana? ¿Quién va a ganar la liga?

3 Se ha realizado un estudio sobre el uso de una determinada configuración de computador en empresas. Se estableció una clasificación de empresas de 1 a 24 en función de aspectos sectoriales, económicos y plantilla; se hizo una encuesta con el objeto de determinar, en función del tipo de empresa, cuántas utilizaban la citada configuración, con los resultados siguientes:

Tipo	Nº Casos	Tipo	Nº Casos	Tipo	Nº Casos
1	3	***9***	100	***17***	56
2	3	***10***	64	***18***	55
3	14	***11***	40	***19***	44
4	26	***12***	15	***20***	21
5	32	***13***	35	***21***	15
6	13	***14***	25	***22***	4
7	24	***25***	24	***23***	4
8	42	***26***	19	***24***	4

a) ¿Cuál es el tamaño de la muestra? Calcular la media, mediana, desviación típica y cuartiles.
b) Construir el gráfico caja y comentar los resultados (dispersión, valores extremos, simetría…).
c) Construye de nuevo el gráfico generando previamente 4 intervalos. Observando el gráfico contesta: ¿Cuáles son los valores para el primer, segundo y tercer cuartil? ¿Coinciden con los del primer gráfico? Explica razonadamente el resultado.