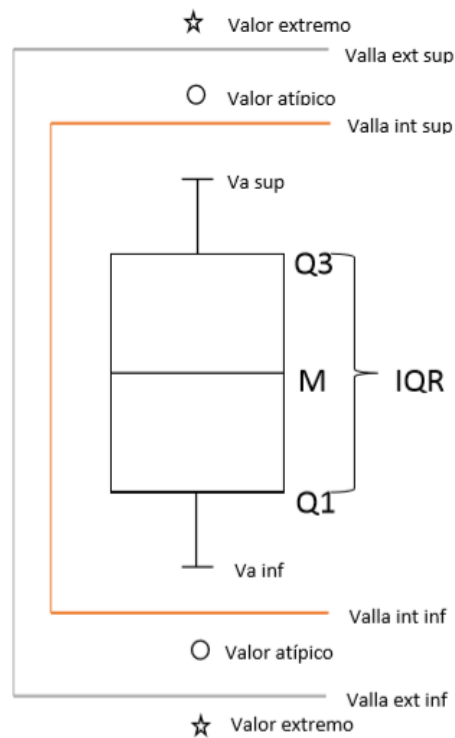


Segundo Examen de Estadística

Práctica 5:

Construcción Gráficos-Caja



- Siendo n el tamaño de la muestra, se pueden calcular las posiciones de:
 - $Q1 = (n+1)/4$
 - $M \text{ (Mediana)} = (n+1)/2$
 - $Q3 = 3(n+1)/4$
 - Para que las posiciones sean enteras, usaremos muestras de tamaño 3, 7, 11, 15,...
- IQR (intervalo intercuartílico) $Q3 - Q1$
- $V_{ext \text{ sup}}$ (valla exterior superior) $Q3 + (3 * IQR)$
- $V_{int \text{ sup}}$ (valla interior superior) $Q3 + (1,5 * IQR)$
- $V_{int \text{ inf}}$ (valla interior inferior) $Q1 - (1,5 * IQR)$
- $V_{ext \text{ inf}}$ (valla exterior inferior) $Q1 - (3 * IQR)$

- **Va sup** (valor adyacente superior) $\max x \leq V_{int\ sup}$.
Máximo dato entre Q3 y dicha valla. El límite es el valor de la valla y puede coincidir con él.
- **Va inf** (valor adyacente inferior) $\min x \geq V_{int\ inf}$.
Mínimo dato entre Q1 y dicha valla. El límite es el valor de la valla y puede coincidir con él.
- **Valores atípicos:** Entre las dos vallas superiores ($>V_{int\ sup}$ y $\leq V_{ext\ sup}$) ó entre las dos vallas inferiores ($\geq V_{ext\ inf}$ y $< V_{int\ inf}$)
- **Valores extremos:** Más alejados de la valla exterior superior ($>V_{ext\ sup}$) ó de la valla exterior inferior ($<V_{ext\ inf}$)

Ejemplo 3:

Una empresa fabricante de componentes mecánicos tiene dos líneas de fabricación (L1 y L2). Se muestra el diagrama Box-Whisker codificado según la línea en que se fabricó cada pieza. Teniendo cuenta que las especificaciones de fabricación dicen que la longitud de las piezas debe estar en el intervalo 100 ± 1 cms, indicar cuál sería la línea más adecuada para fabricar las piezas, justificando la respuesta.

Los diagramas de caja son muy útiles al hacer comparaciones gráficas entre conjuntos de datos.

Para ello debemos fijarnos en tres cosas básicas:

"Normalidad, Amplitud de caja y presencia de datos anómalos".

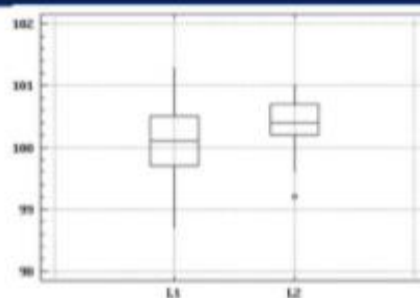


Imagen 4, Diagrama Box-Whisker

Fuente: Romero, R y Zúñiga, L.R., (1993)

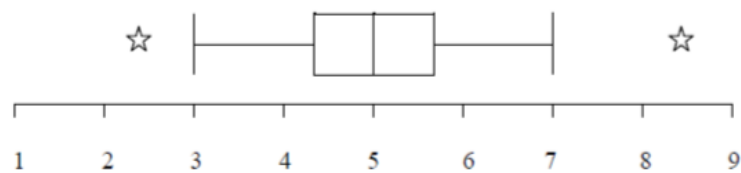
La línea L1 fabrica mayor proporción de piezas con dimensión cercana al nominal, pero produce unas pocas piezas fuera de especificación. La línea L2 produce las piezas con menor dispersión, todas dentro de la especificación, pero con una dimensión sistemáticamente mayor que el nominal. Convendría usar la línea L2 pero intentar centrar el proceso de fabricación alrededor de la media deseada.

INTERPRETACIÓN DE UN GRÁFICO-CAJA

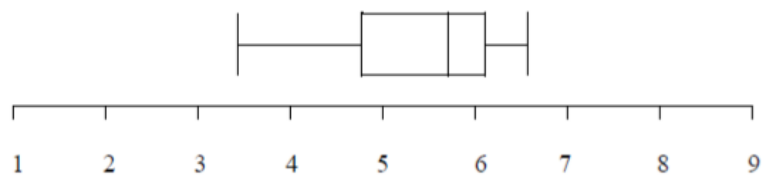
A la hora de interpretar un gráfico-caja tenemos que fijarnos en su simetría, que indica la forma del conjunto de datos, lo cual implica observar dónde se concentra la información. Para el estudio de la forma de una distribución, también se usan los términos sesgo o asimetría.

Una distribución puede ser:

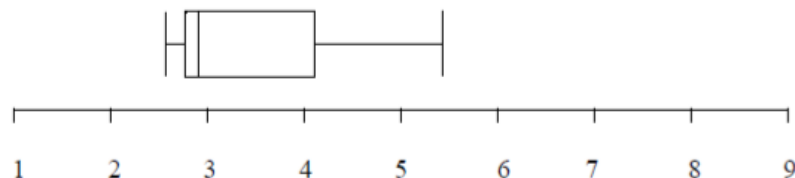
- **Simétrica:** en este tipo de distribuciones la media, la moda y la mediana coinciden y se encuentran ubicadas en el centro de la caja, y los datos se distribuyen de igual forma a ambos lados de estas medidas.



- **Asimétrica positiva o sesgada a la derecha:** los datos tienden a concentrarse hacia la parte inferior de la distribución, la mediana está más cerca del lado izquierdo de la caja (valores inferiores) y el bigote derecho es más largo. La media suele ser mayor que la mediana en estos casos. Los valores menores están más concentrados mientras que los valores mayores están más dispersos.



- **Asimétrica negativa o sesgada a la izquierda:** los datos tienden a concentrarse hacia la parte superior de la distribución, la mediana está más cerca del lado derecho de la caja (valores superiores) y el bigote izquierdo es más largo. La media suele ser menor que la mediana en estos casos. Los valores mayores están más concentrados mientras que los valores menores están más dispersos.



Práctica 6:

Ejercicio:

Correlaciones			
		Horas de estudio	Calificación
X	Correlación de Pearson	1	,945 [*]
	Sig. (bilateral)		,015
	Suma de cuadrados y productos cruzados	319,200	50,100
	Covarianza	79,800	12,525
	N	5	5
Y	Correlación de Pearson	,945 [*]	1
	Sig. (bilateral)	,015	
	Suma de cuadrados y productos cruzados	50,100	8,800
	Covarianza	12,525	2,200
	N	5	5

*. La correlación es significativa al nivel 0,05 (bilateral).

❑ **Covarianza:** $Cov(X,Y)=12.525$

❑ **Correlación de Pearson:**

$R=0.945 \rightarrow$ Fuerte dependencia lineal creciente, el ajuste lineal es apropiado.

❑ **Coefficiente de determinación:** $R^2=0.893 \rightarrow$ La variable horas de estudio explica el 89.3% de la variabilidad de la variable calificación.

Ejercicio:

Resumen del modelo y estimaciones de los parámetros

Variable dependiente: Calificación

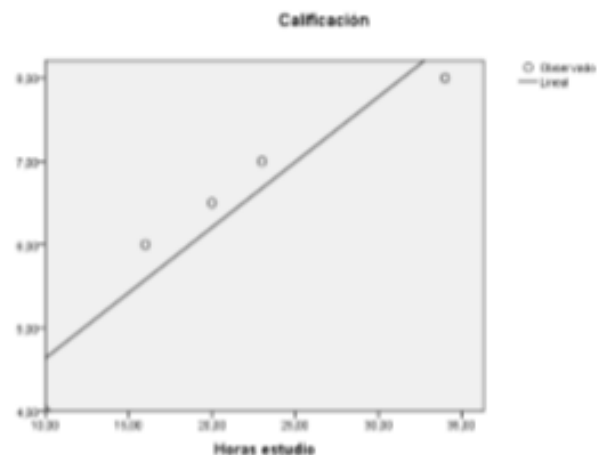
Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,894	25,188	1	3	,015	3,067	,157

La variable independiente es Horas de estudio.

Recta de regresión



$$Y=0.157X+3.067$$



Ejercicio: Si una persona ha estudiado 15 horas , ¿cuánto cabe esperar que haya sacado en el examen? **5.422**

Resumen del modelo y estimaciones de los parámetros

Variable dependiente: Calificación

Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	.894	25,188	1	3	.015	3,067	.157

La variable independiente es: horas de estudio.



Recta de regresión

$$Y = 0.157X + 3.067$$



$$Y = 0.157 \cdot 15 + 3.067 = 5.422$$

