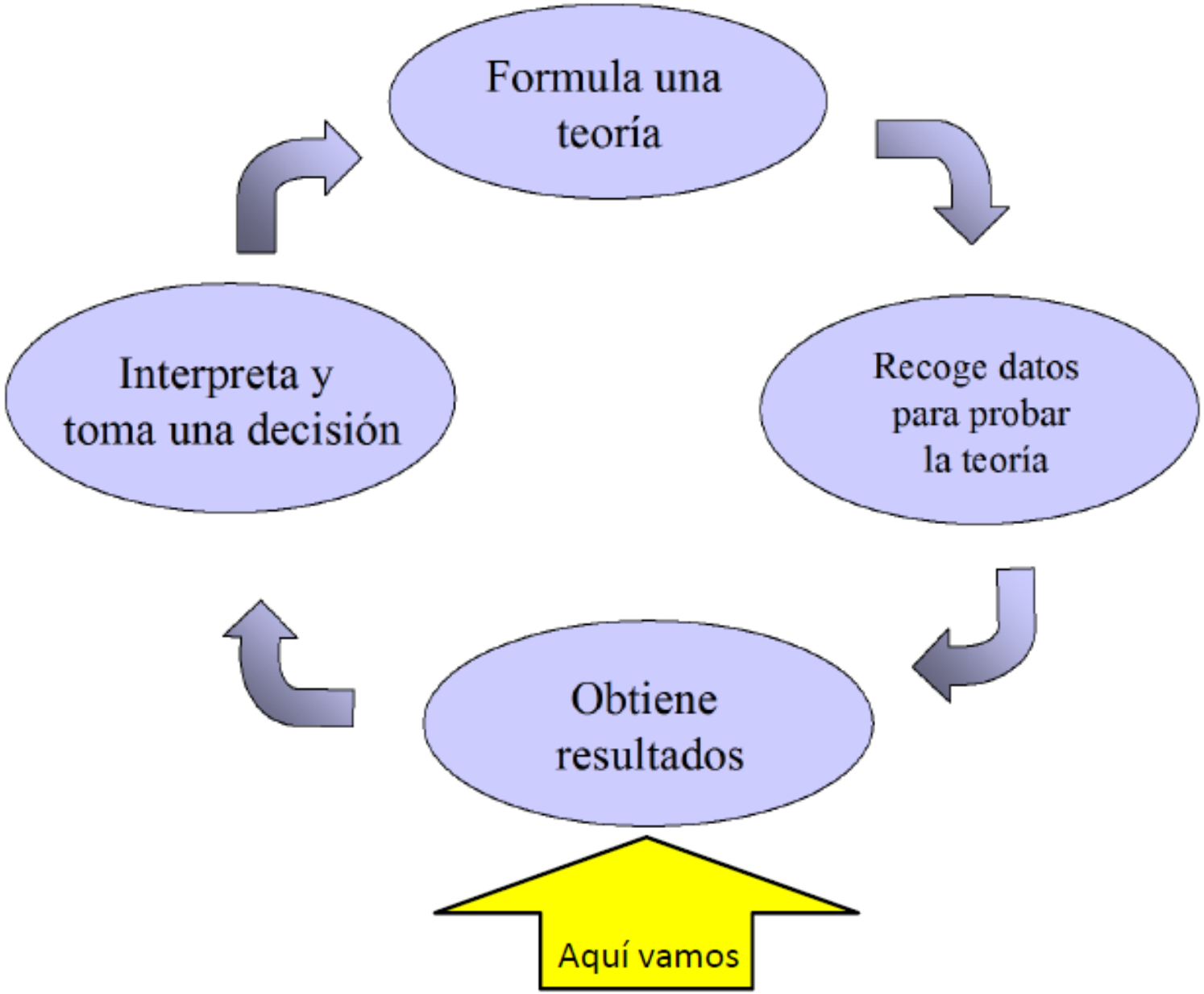
UNIVERSIDAD DE TALCA

INSTITUTO DE MATEMÁTICA Y FIsiCA

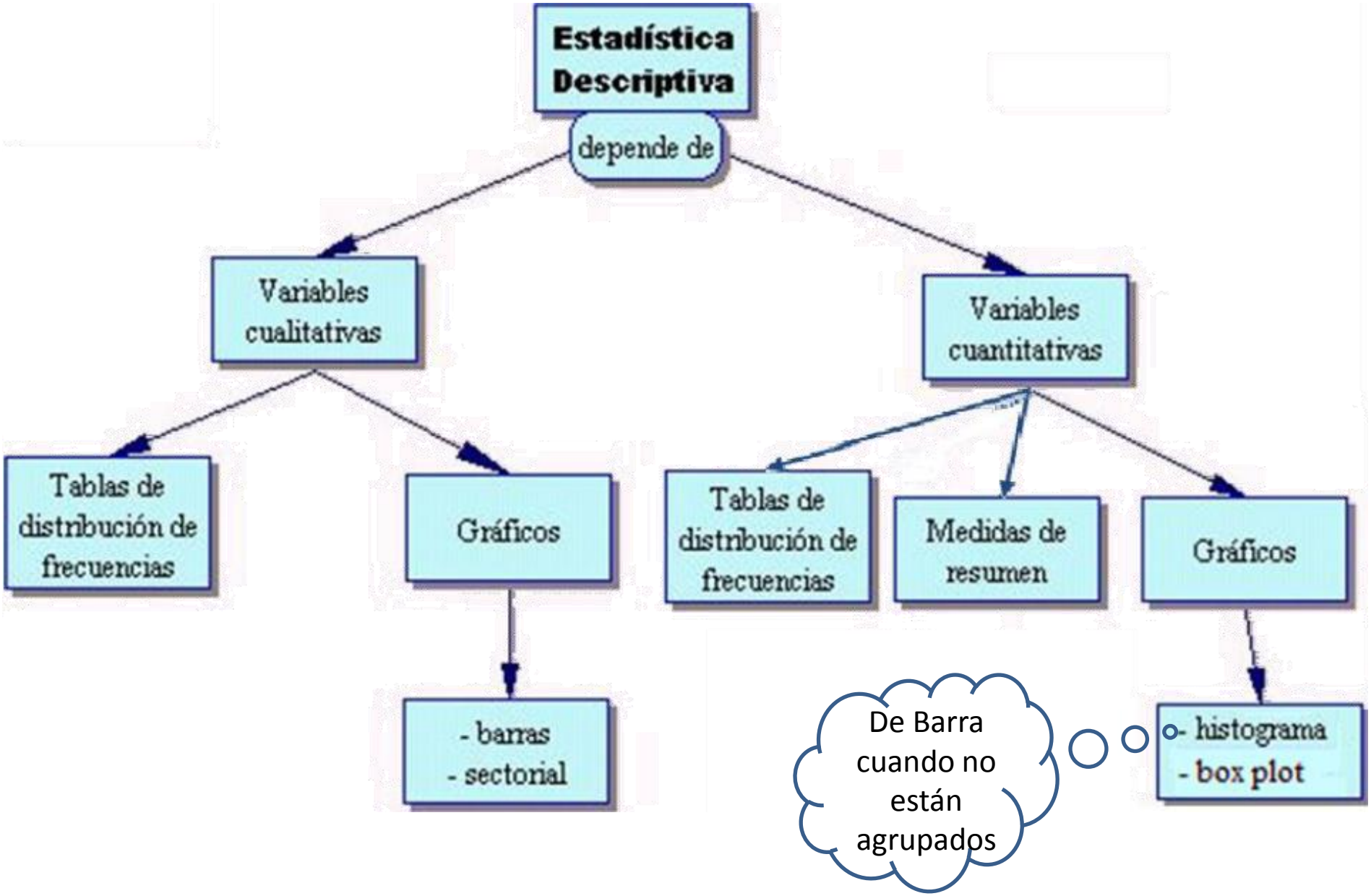
Análisis Exploratorio de Datos
Resumen gráfico y numérico

DOCENTE Gloria Correa Beltrán

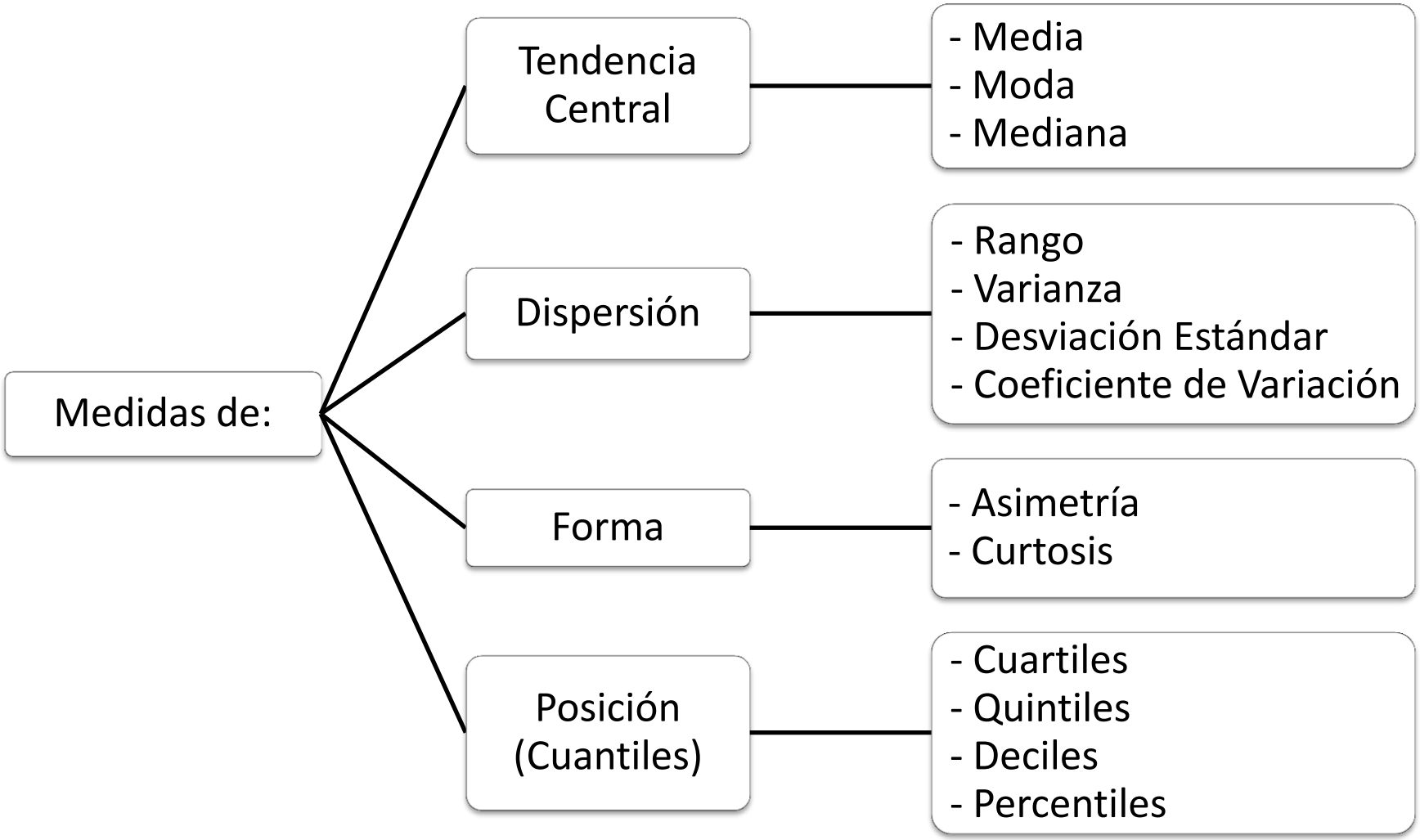
Etapas del Método Científico



Pasos a seguir en el Análisis Exploratorio de Datos



Medidas Descriptivas Numéricas



Medidas de Tendencia Central

Media Aritmética

- Se calcula sumando el valor de todos los datos y dividiéndolos por el número total de ellos.
- Es muy sensible a los valores extremos.

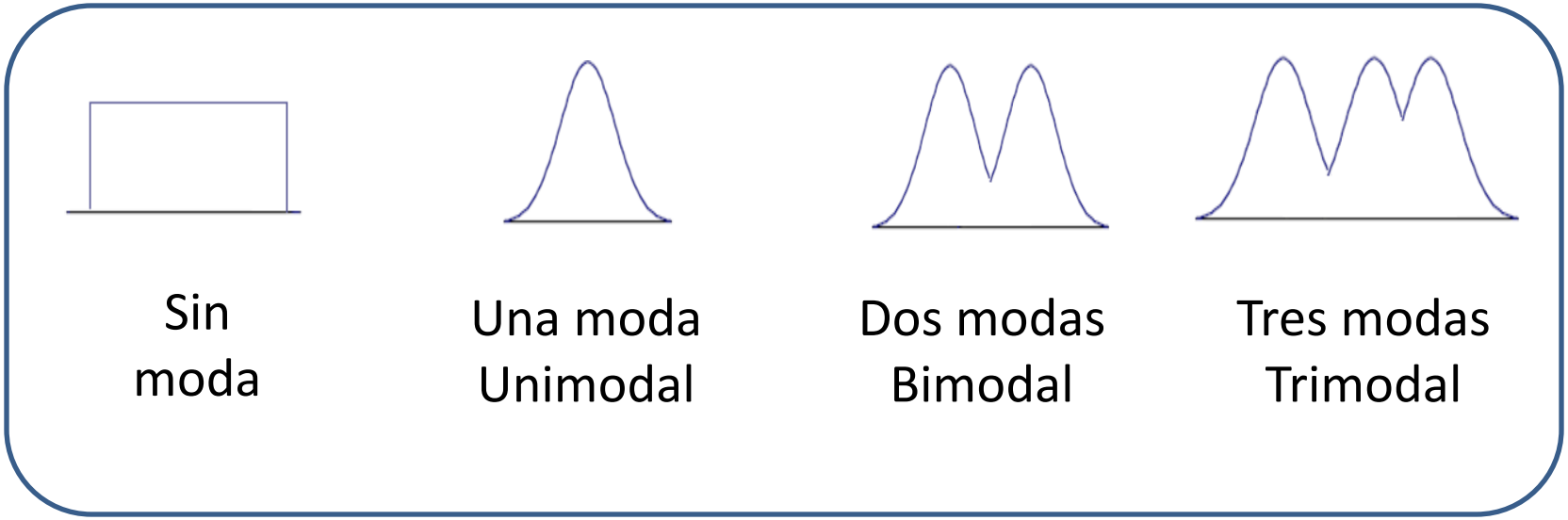
Moda

- Se define como el valor que se presenta con más frecuencia.
- No es sensible a valores extremos.
- Cuando es obtenida a partir de una tabla con datos sin agrupar, la moda es el valor con más alta frecuencia. Y cuando es obtenida a partir de una tabla con datos agrupados, la moda es la marca de clase con más alta frecuencia.

Medidas de Tendencia Central

Moda

Puede no ser un estadígrafo único, en una distribución de frecuencias podría existir más de una moda (bimodal, trimodal), también es posible que alguna distribución no tenga moda.



Medidas de Tendencia Central

Mediana

- Divide la distribución en dos partes con igual cantidad de datos.
- Si los datos se ordenan en forma creciente o decreciente, la mediana es el valor que se encuentra en el medio de la distribución.
- Si el número de datos es impar entonces la mediana coincide con el dato que se encuentra en la mitad.
- Si el número de observaciones es par, habrá dos observaciones centrales en ese caso la mediana será el promedio de ellas.
- La mediana no es sensible a valores extremos, esta medida depende del número de observaciones y no de la magnitud de ellas.

Dispersión

Variabilidad de los datos ≠ Homogeneidad de los datos

- Alejamiento o concentración de los datos con respecto a un valor central (promedio).

- Una dispersión pequeña indica que es poco el alejamiento de los datos con respecto al valor central.

Poco disperso

Muy disperso

Medidas de Dispersión

Rango o Recorrido

$$R = x_{\max} - x_{\min}$$

- Se obtiene restando el dato mayor con el menor.
- Es la Medida de Dispersión más elemental.
- Proporciona información imprecisa, pues sólo considera los valores extremos de la distribución.
- Es independiente de las Medidas de Tendencia Central.

Medidas de Dispersión

Varianza

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n}$$

- Promedio de las desviaciones (distancia, resta) al cuadrado de los datos con respecto a la media aritmética.
- Unidad de medida de los datos al cuadrado.

Desviación Estándar

$$S = \sqrt{S^2}$$

- Raíz cuadrada positiva de la varianza.
- Misma unidad de medida de los datos.
- Más fácil de interpretar.

Medidas de Dispersión

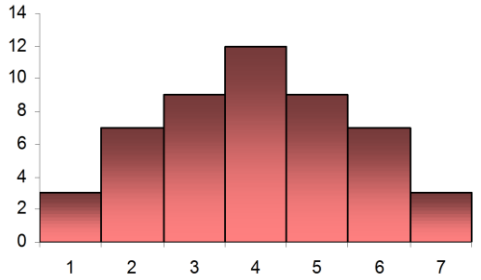
Coeficiente de Variación

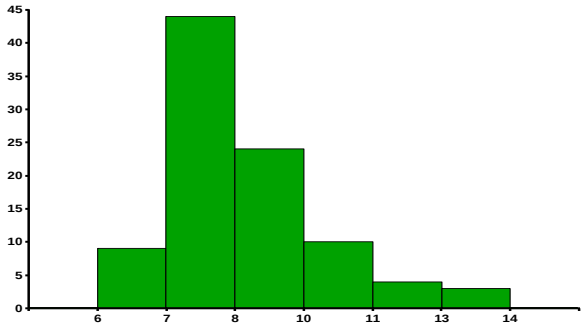
$$CV = \frac{S}{\overline{x}} \cdot 100$$

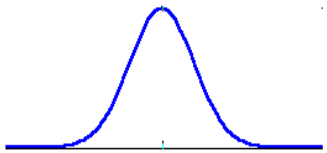
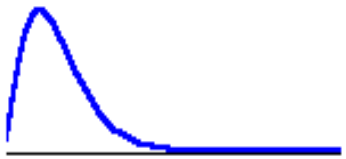
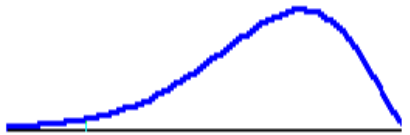
- Medida de dispersión adimensional, que entrega el resultado en porcentaje.
- Permite comparar dos distribuciones independiente de sus unidades de medida.
- Dados dos conjuntos, el menos disperso es aquel que tenga un Coeficiente de Variación menor.
- Un Coeficiente de Variación mayor al 30% indica que los datos están muy dispersos, no son homogéneos.

Medidas de Forma
Asimetría o **Sesgo**

- Falta de simetría de una distribución de frecuencias con respecto a una distribución simétrica unimodal.
- Una distribución es asimétrica si el lado derecho e izquierdo son diferentes con respecto a un valor central.



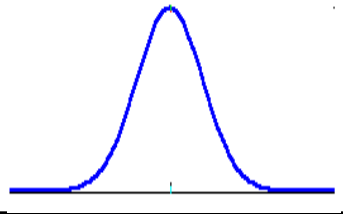
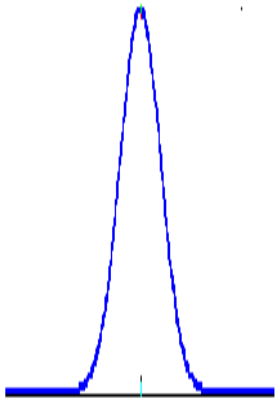
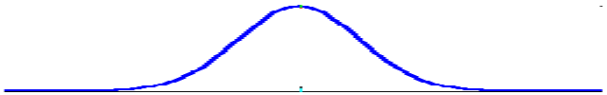


Simétrica (sin sesgo)	Asimétrica a la derecha (sesgo positivo)	Asimétrica a la izquierda (sesgo negativo)
$= 0$	> 0	< 0
		

Medidas de Forma

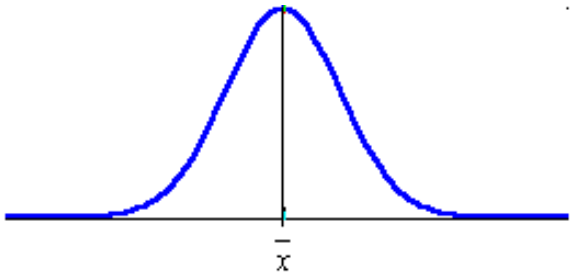
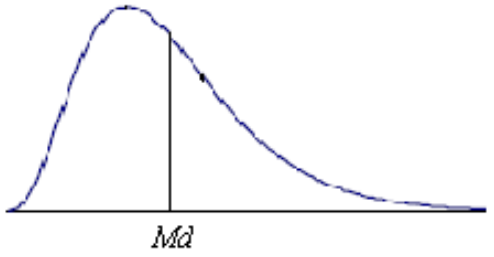
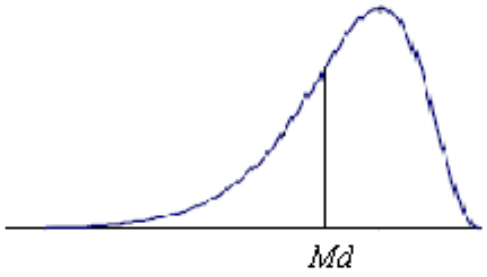
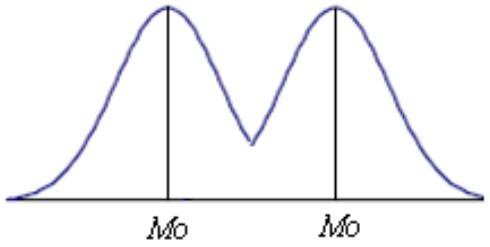
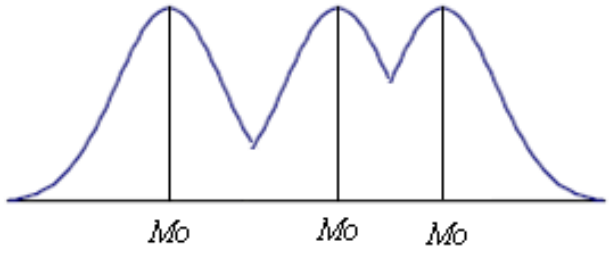
Curtosis

Grado de apuntamiento de una distribución de frecuencias con respecto a una distribución simétrica unimodal de forma acampanada.

Mesocúrtica (acampanada)	Leptocúrtica (espigada)	Platicúrtica (achatada)
$= 0$	> 0	< 0
		

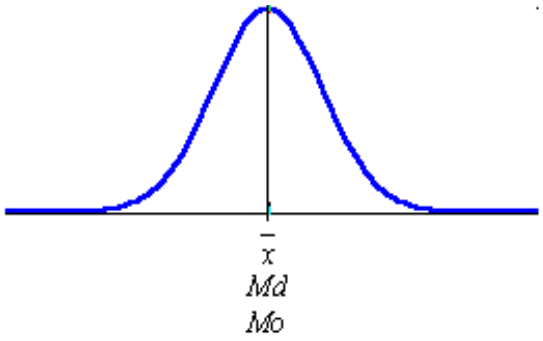
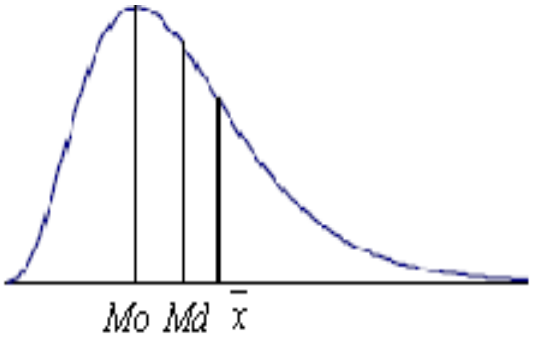
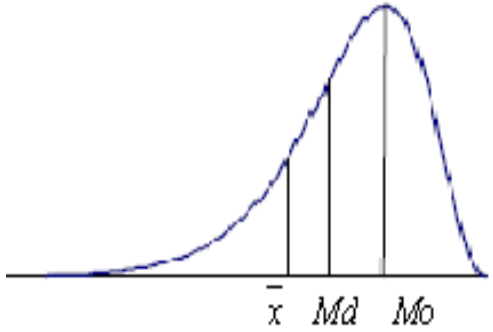
Medidas de Tendencia Central
Medidas numéricas representativas de una distribución de frecuencias

Si la distribución de frecuencias es:

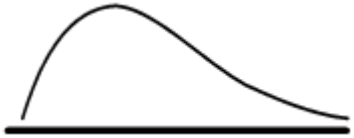
Simétrica	**Asimétrica**	**Sinusoidal**
La **Media** es el valor más representativo de la distribución por estar equidistante de los extremos.	La **Mediana** es mejor representante que la media.	La **Moda** es más representativa que la media y la mediana.
	 	 

Orden de Medidas de Tendencia Central en una distribución de frecuencias

Si la distribución de frecuencias es:

Simétrica	Asimétrica a la derecha	Asimétrica a la izquierda
$\bar{x} = Md = Mo$	$Mo < Md < \bar{x}$	$\bar{x} < Md < Mo$
		

Formas de Distribuciones

Simétrica, Acampanada, Unimodal  Ej. Puntajes de la PSU	**Bimodal**  Ej. Estatura de la población de hombres y mujeres
Asimétrica a la derecha  Ej. Sueldos de una industria	**Asimétrica a la izquierda**  Ej. Notas de Bioestadística

¿Qué forma tiene la distribución de la estatura?

El entrenador antes del partido pregunta a sus jugadores:

¿qué hacemos

le mandamos a los rivales la estatura promedio del equipo

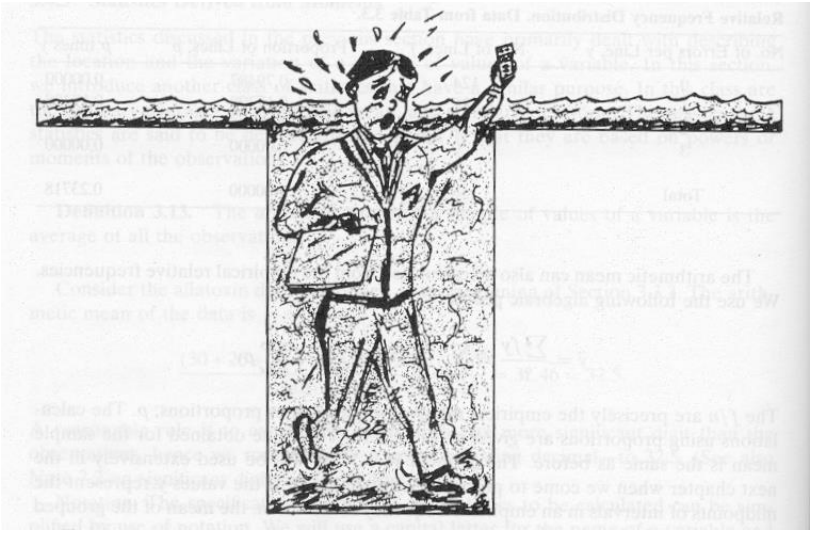
o

los tranquilizamos con la estatura mediana?.

Un alumno de bioestadística olvidó que la variabilidad es importante

…

aquí lo vemos ahogándose en un río con una profundidad promedio de 30 centímetros.

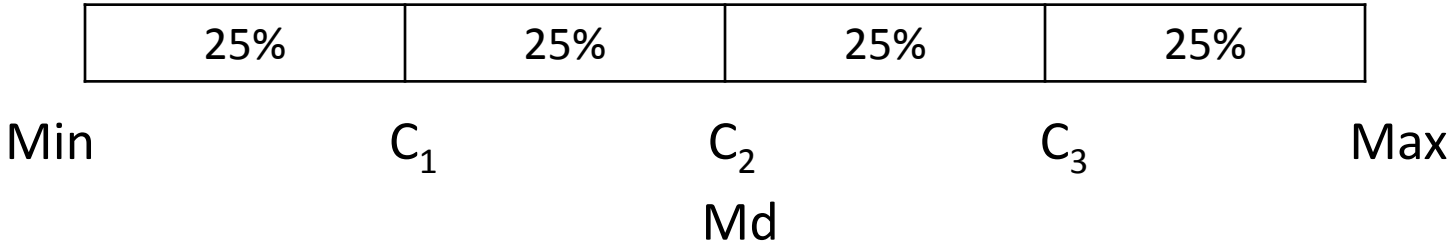


Medidas de Posición o Cuantiles

Valores que dividen a la distribución en cuatro, cinco, diez o cien partes iguales.

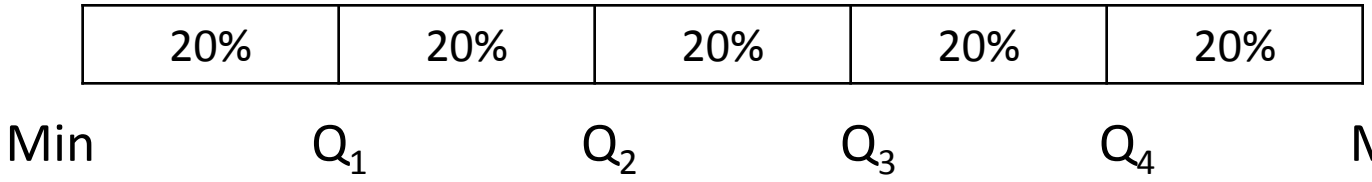
Cuartiles

Valores que dividen a una distribución de frecuencias en cuatro partes iguales, mediante el primer, segundo y tercer cuartil.



Quintiles

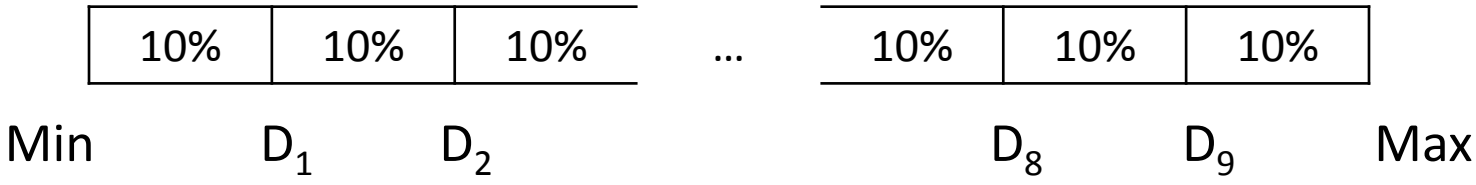
Valores que dividen a una distribución de frecuencias en cinco partes iguales, mediante el primer, segundo, tercer y cuarto quintil.



Medidas de Posición o Cuantiles

Deciles

Valores que dividen a una distribución de frecuencias en diez partes iguales, mediante el primer, segundo, … y noveno decil.



Percentil

Valores que dividen a una distribución de frecuencias en cien partes iguales, mediante el primer, segundo, … y 99^{avo} percentil.

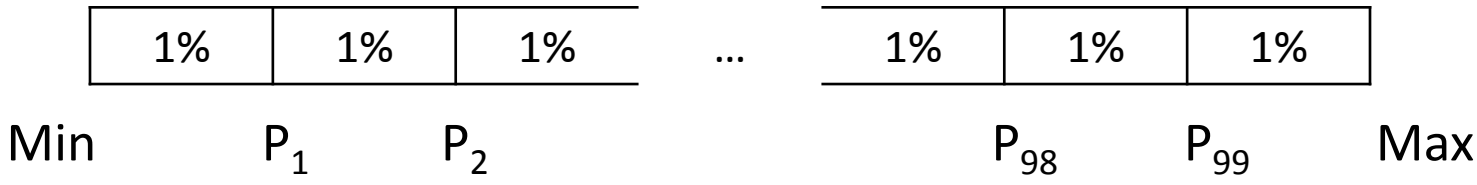


GRAFICO DE CAJA

Box Plot

El interior de la caja encierra el 50% central de los casos. La longitud de la caja da una idea de la variabilidad de los datos.

Sus límites son el 25 % superior e inferior de los casos.

El límite inferior de la caja representa el cuartil 1 o percentil 25.

El límite superior es el cuartil 3 ó percentil 75.

La longitud de la caja es el rango intercuartil:

$$Q = Q_3 - Q_1$$

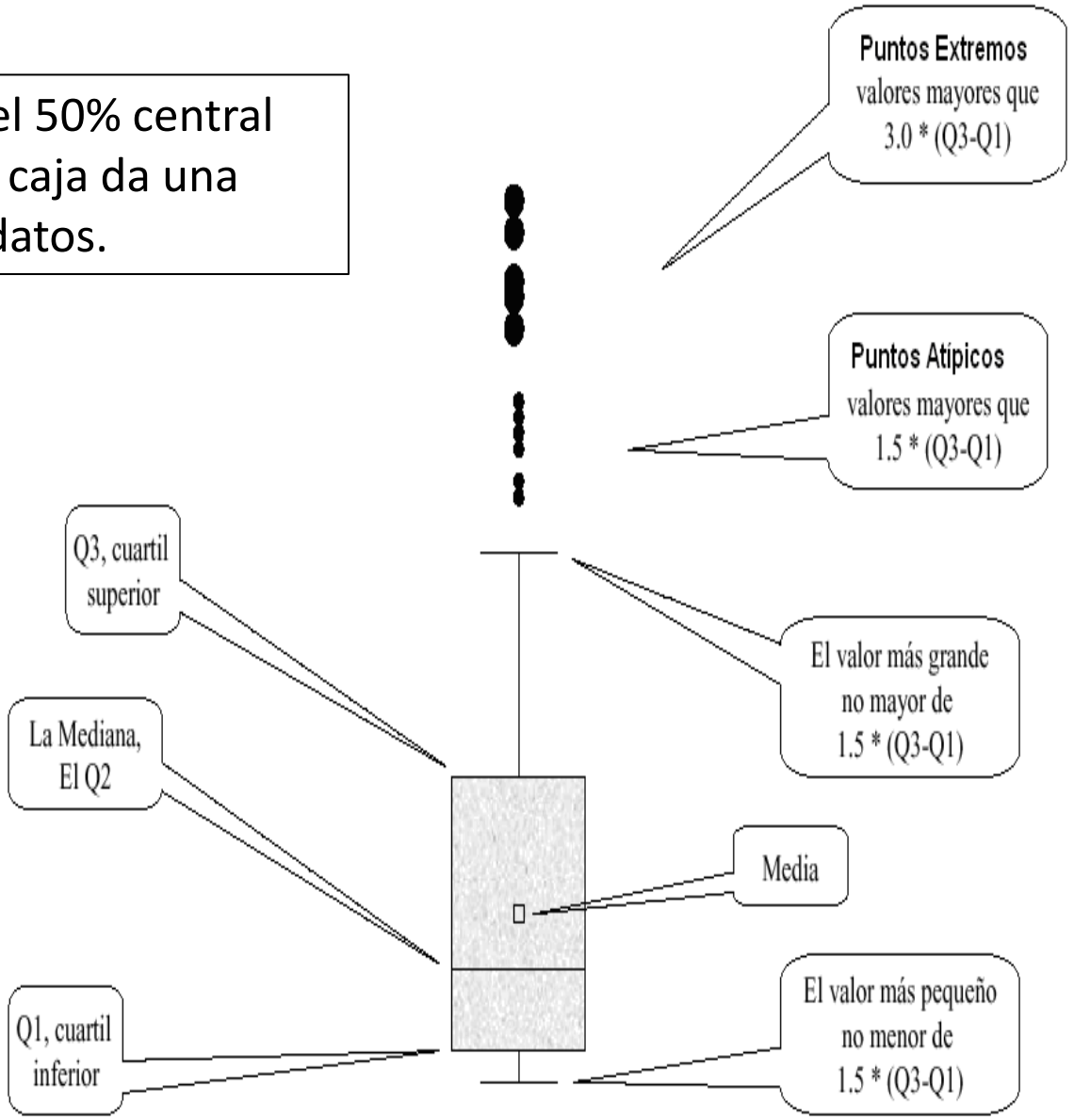


GRAFICO DE CAJA

Box Plot

El pequeño cuadradito dentro de la caja representa la media y la línea que divide a la caja en dos partes es la mediana (cuartil 2 ó percentil 50).

La ubicación de la media y la mediana da una idea de la tendencia central de los datos.

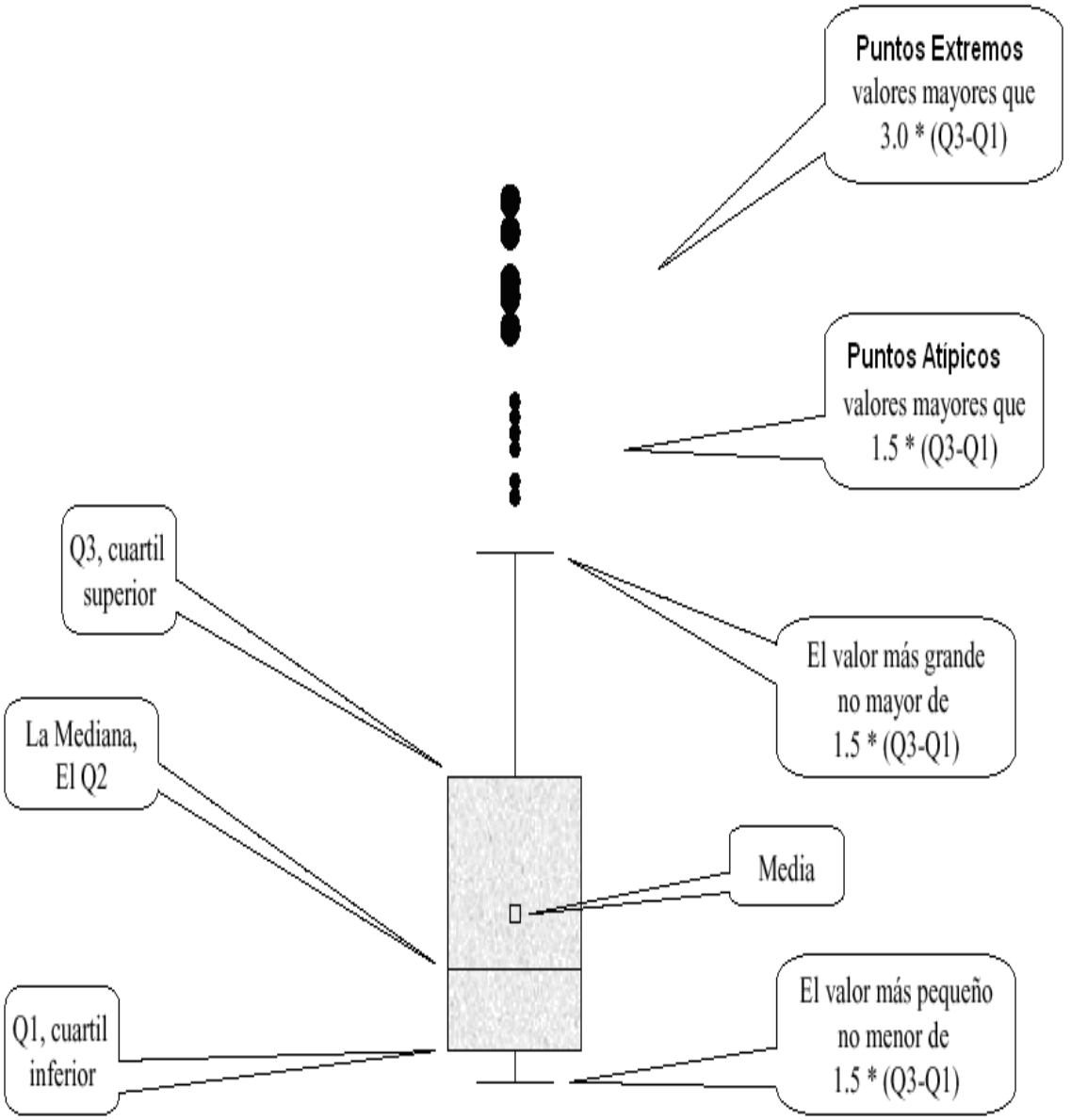


GRAFICO DE CAJA

Box Plot

Desde los extremos de la caja se trazan líneas hasta los respectivos valores adyacentes. A estas líneas se les llama “antenas”.

Se llaman puntos atípicos u outliers a aquellos datos que se encuentran fuera de las barreras internas y dentro de las barreras externas.

Se llaman puntos extremos a aquellos puntos ubicados fuera de las barreras externas.

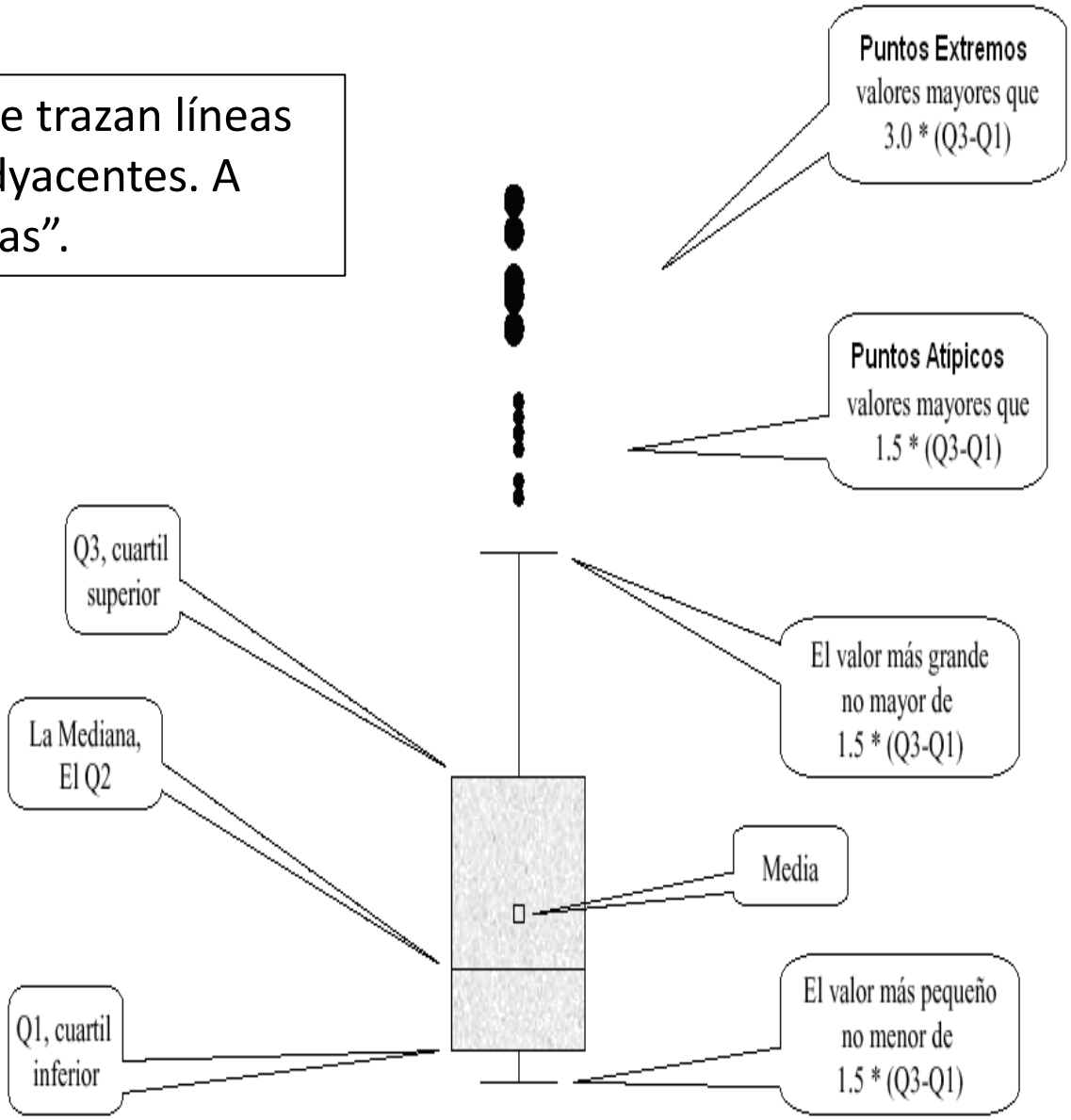


GRAFICO DE CAJA

Box Plot

A través de un gráfico caja se puede identificar el tipo de asimetría de una distribución de frecuencias unimodal:

- Si la posición de la mediana se encuentra en la mitad de la caja y las antenas tienen la misma longitud, la distribución es simétrica.

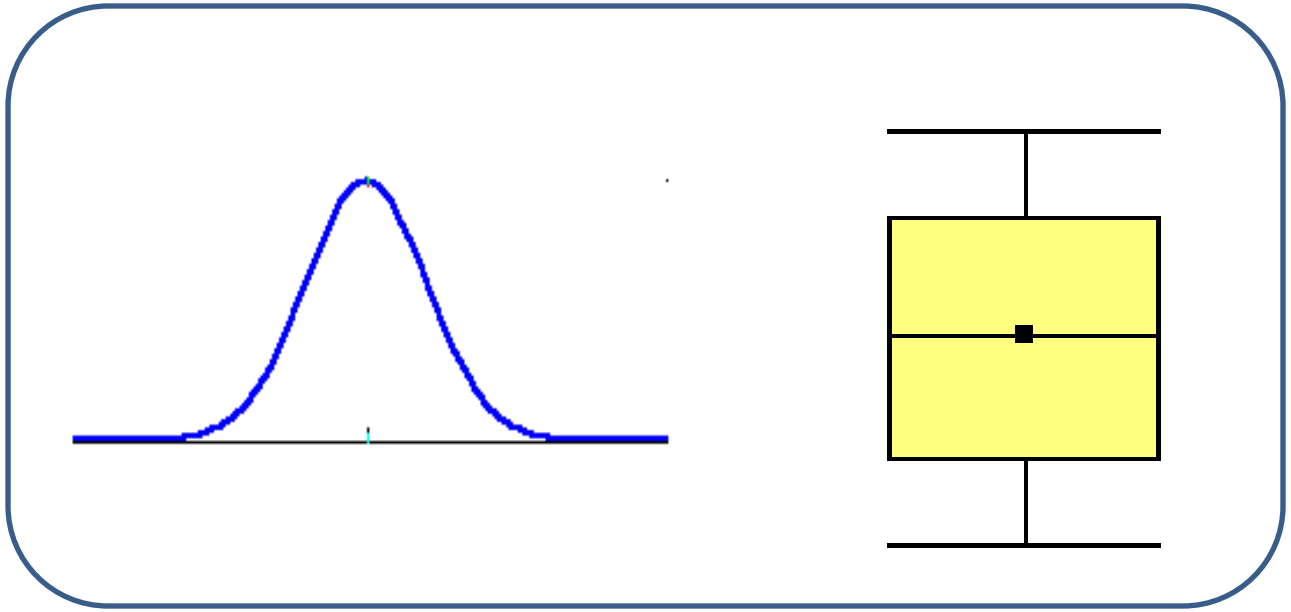


GRAFICO DE CAJA

Box Plot

- Si la posición de la mediana se encuentra ubicada más cerca del primer cuartil y la antena superior es de mayor longitud que la antena inferior, la distribución presenta sesgo positivo.

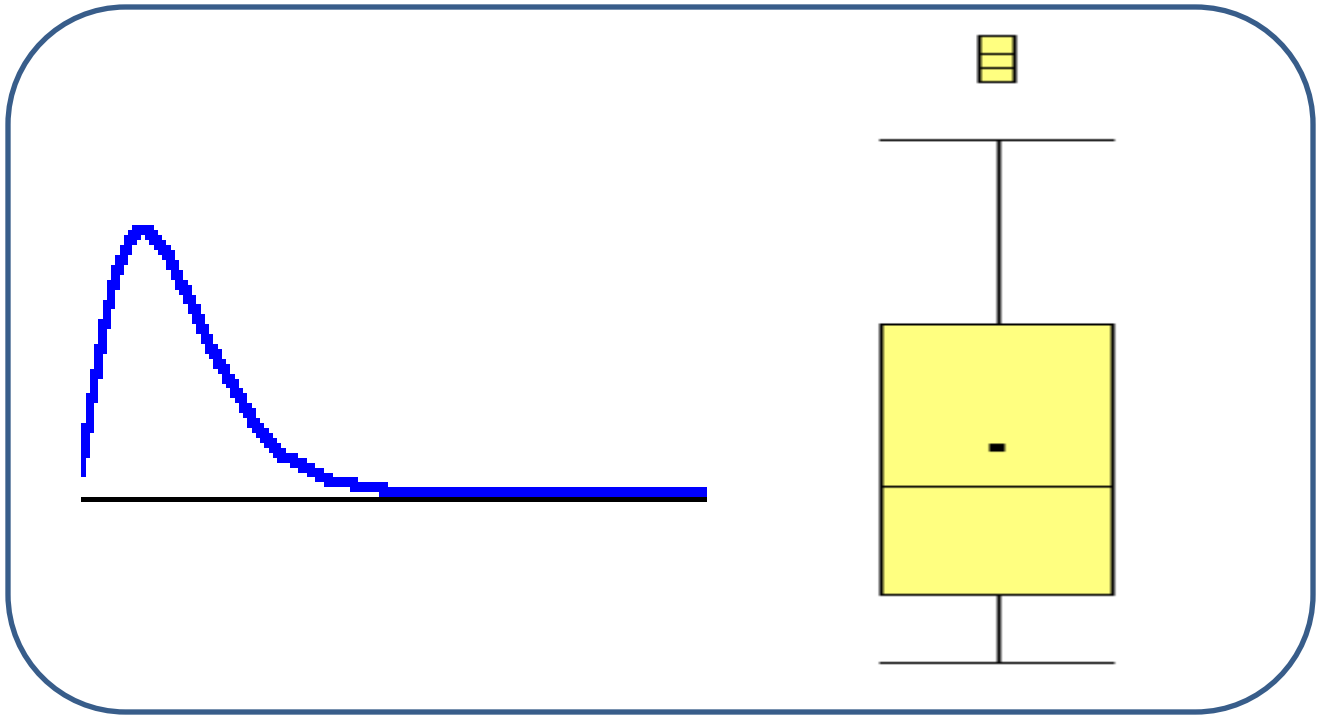


GRAFICO DE CAJA

Box Plot

- Si la posición de la mediana se encuentra ubicada más cerca del tercer cuartil y la antena superior es de menor longitud que la antena inferior, la distribución presenta sesgo negativo.

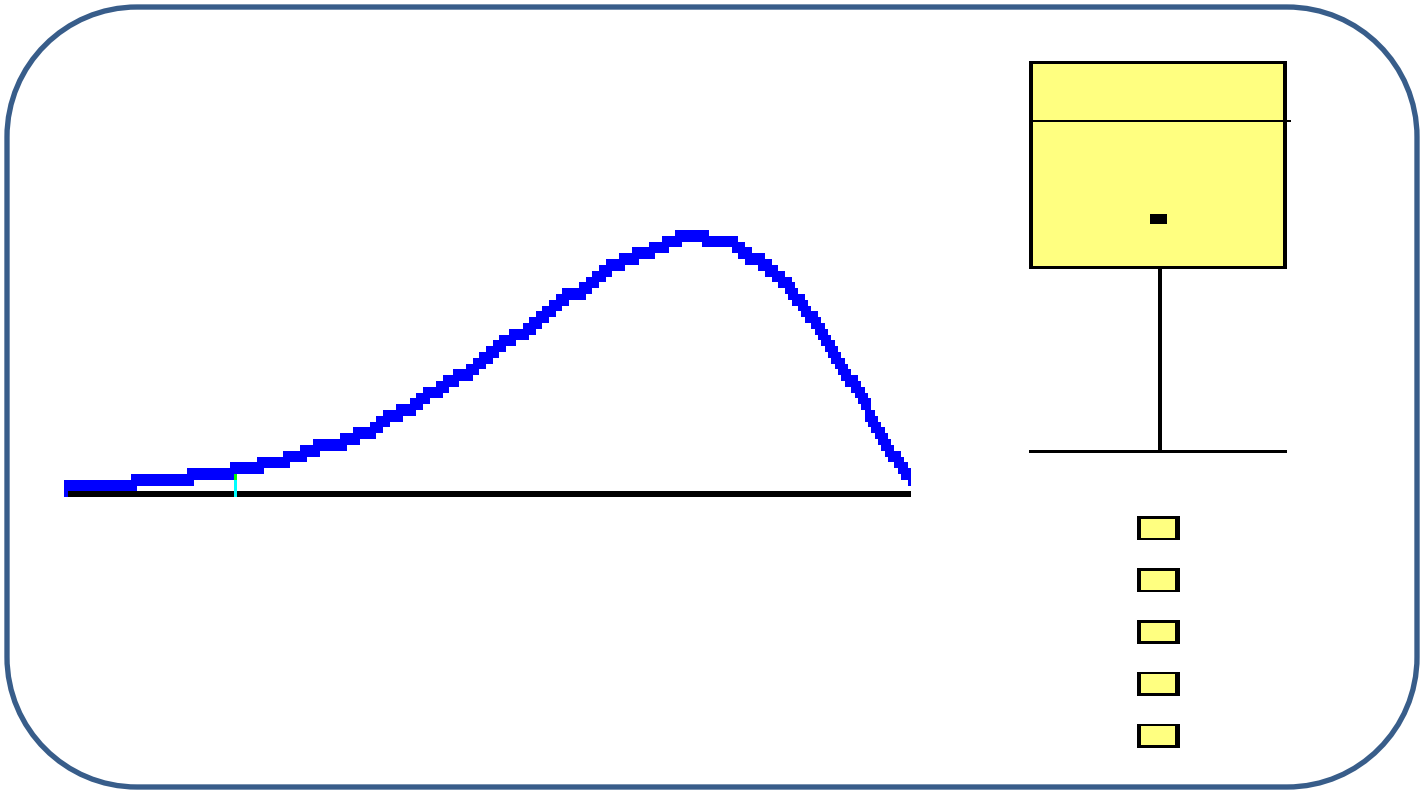
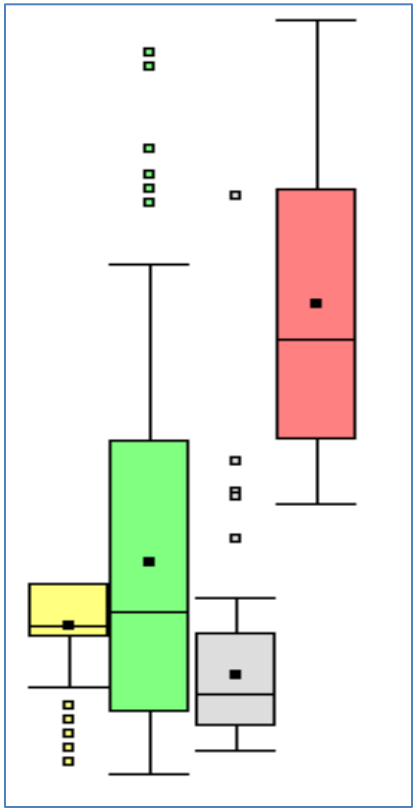


GRAFICO DE CAJA

Box Plot

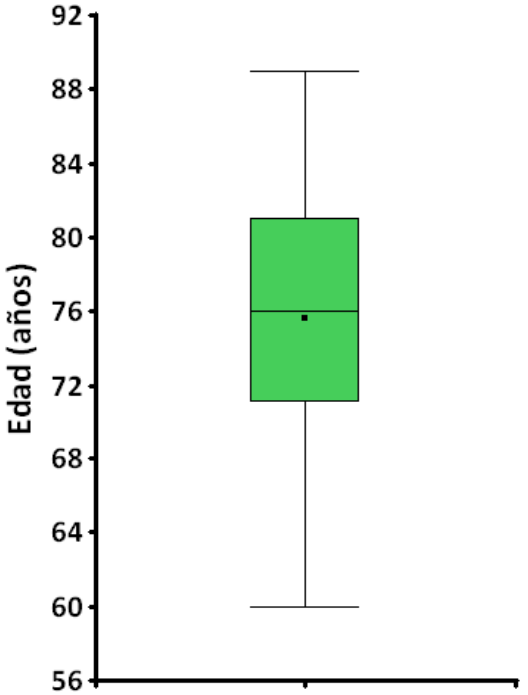
Este tipo de representación es especialmente útil para comparar:

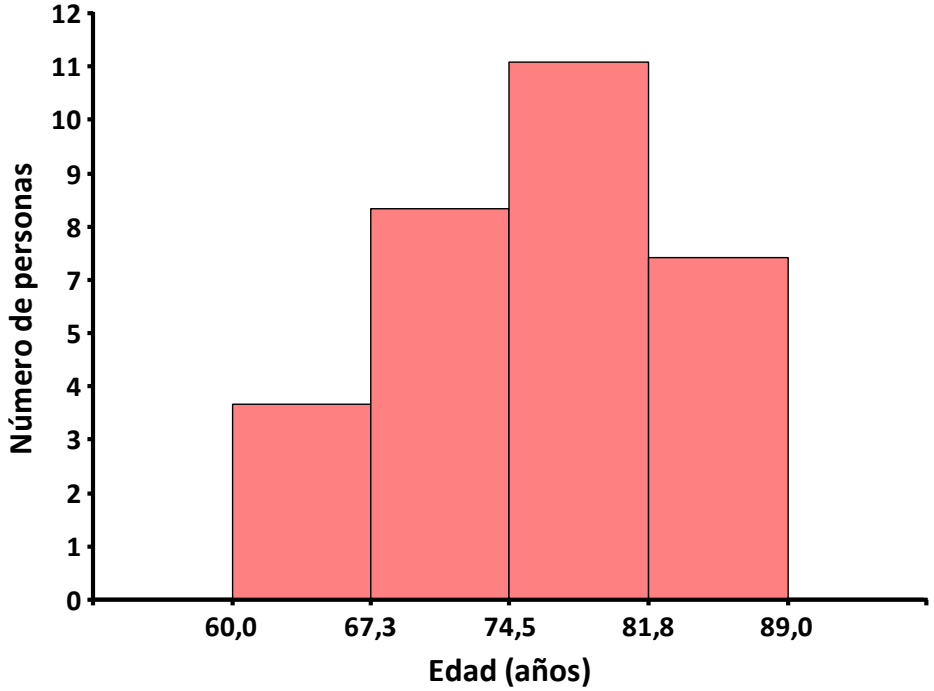
- La distribución de los datos de una misma variable en varios grupos.
- Una misma variable medida en momentos diferentes.



Volvamos al ejemplo anterior

Edad	Marca de clase	Número de personas	Porcentaje de personas	Número de personas acumulado	Porcentaje de personas acumulado
[60,0 ; 67,3]	63,65	4	13,3	4	13,3
(67,3 ; 74,5]	70,90	8	26,7	12	40,0
(74,5 ; 81,8]	78,15	11	36,7	23	76,7
(81,8 ; 89,0]	85,40	7	23,3	30	100,0
Total		30	100,0		





Estadístico	Edad
Minimo	60
Cuartil 1	71
Media	75,67
Mediana	76
Cuartil 3	81
Máximo	89
Asimetría	-0,12
Kurtosis	-0,69
Desviación Estándar	7,93
Coeficiente de Variación	10,48

GRAFICO DE CAJA

Construcción del Box Plot

- Se traza un rectángulo cuyos extremos se ubican en el primer y tercer cuartil.
- En la caja se traza una recta horizontal en el lugar de la mediana.
- En el centro horizontal y a la altura de la media se dibuja un punto.
- Se calcula el rango intercuartil $(C = C_3 - C_1)$.
- Calcular las barreras internas

$$BI_1 = C_1 - 1.5\ C \qquad BI_2 = C_3 + 1.5\ C$$

- Calcular las barreras externas

$$BE_1 = C_1 - 3C \qquad BE_2 = C_3 + 3C$$

GRAFICO DE CAJA

Construcción del Box Plot

- Identifica los puntos adyacentes: Se llaman puntos adyacentes al mínimo y máximo dato que se encuentran dentro de las barreras internas. Desde los extremos de la caja se trazan líneas hasta los respectivos valores adyacentes. A estas líneas se les llama “antenas” o “bigotes”.

- Identifica los puntos atípicos: Se llaman puntos atípicos u outliers a aquellos datos que se encuentran fuera de las barreras internas y dentro de las barreras externas.

- Identifica los puntos extremos: Se llaman puntos extremos a aquellos puntos ubicados fuera de las barreras externas

GRAFICO DE CAJA

Ejercicio 1

Los datos corresponden a las cantidades gastadas en miles de pesos por una muestra aleatoria de clientes de un supermercado.

30 - 50 - 90 - 90 - 90 - 100 - 110 - 110 - 110 - 110 - 120 - 150

$\overline{X} = 96,7$

Md = 105

$C_1 = 90$

$C_3 = 110$

C = 20

BI1 = C1 – 1.5 C
= 90 – 1.5 * 20
= 60

BI2 = C3 + 1.5 C
= 110 + 1.5 * 20
= 140

BE1 = C1 – 3C
= 90 – 3 * 20
= 30

BE2 = C3 + 3C
= 110 + 3 * 20
= 170

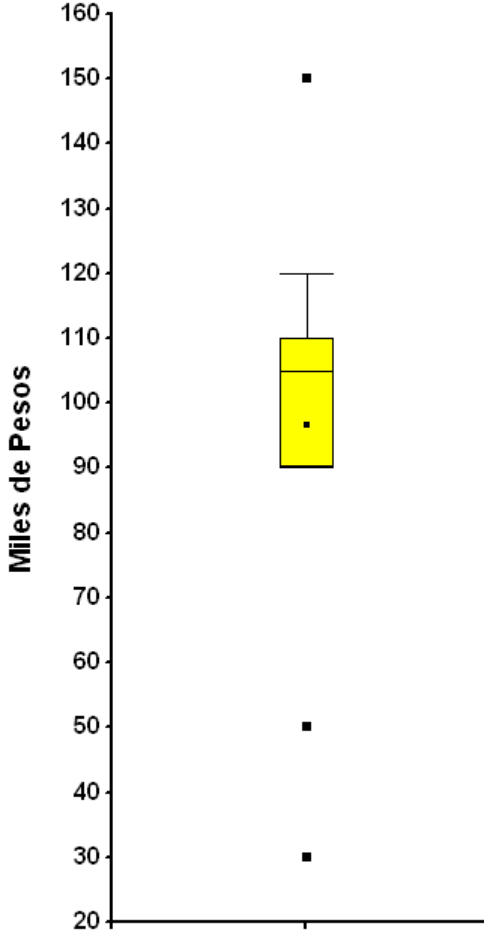
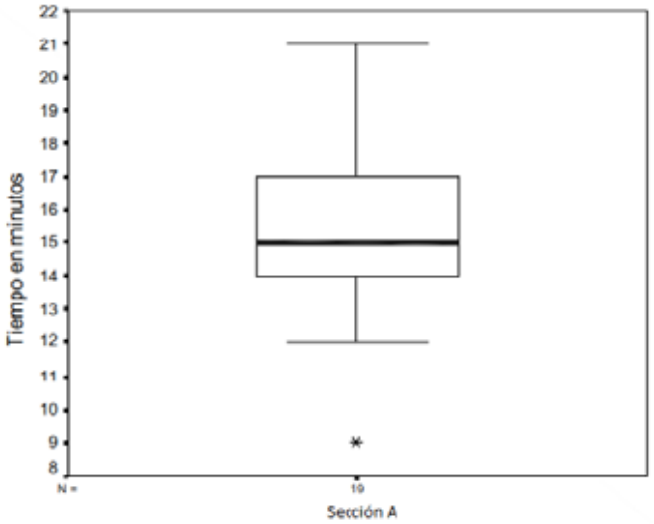


GRAFICO DE CAJA

Ejercicio 2

En una Industria se hizo un listado con los trabajadores que llegaron atrasados un día domingo y para cada uno se registró el tiempo de atraso en minutos. La información fue particionada en dos grupos: Diurno (primer y segundo turno) y Nocturno (tercer turno.

El gráfico muestra los tiempos de atraso de los trabajadores del grupo Diurno:



a. ¿Cuál es el valor aproximado de las medidas de tendencia central y de dispersión del tiempo de tiempos de atraso?

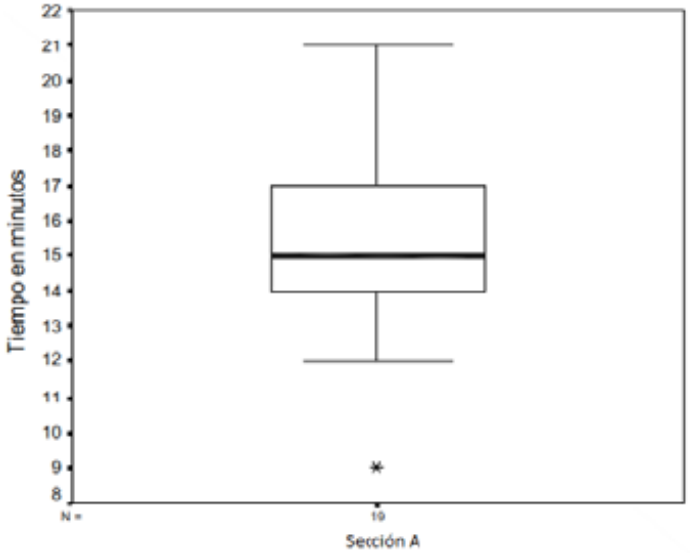
b. ¿Qué porcentaje de trabajadores tiene menos de 14 minutos de atraso? Justifique.

Construya un diagrama de caja para los datos del grupo Nocturno:

10,5 - 11,3 - 11,9 - 12,0 - 12,3 - 12,3 - 12,5 - 12,7 - 13,4 - 13,7
13,8 - 14,2 - 14,8 - 15,1 - 15,3 - 16,7 - 16,8 - 18,8 - 20,8

GRAFICO DE CAJA

Ejercicio 2



a. ¿Cuál es el valor aproximado de las medidas de tendencia central y de dispersión del tiempo de tiempos de atraso?

b. ¿Qué porcentaje de trabajadores tiene menos de 14 minutos de atraso? Justifique.

GRAFICO DE CAJA

Ejercicio 2

Construya un diagrama de caja para los datos del grupo Nocturno:

10,5	11,3	11,9	12,0	12,3	12,3	12,5	12,7	13,4	13,7
13,8	14,2	14,8	15,1	15,3	16,7	16,8	18,8	20,8	

$\overline{X} = 14,2$

Md = 13,7

$C_1 = 12,3$

$C_3 = 15,3$

C = 3,0

BI1 = C1 – 1.5 C

= 12,3 – 1.5 * 3

= 7,8

BI2 = C3 + 1.5 C

= 15,3 + 1.5 * 3

= 19,8

BE1 = C1 – 3C

= 12,3 – 3 * 3

= 3,3

BE2 = C3 + 3C

= 15,3 + 3 * 3

= 24,3

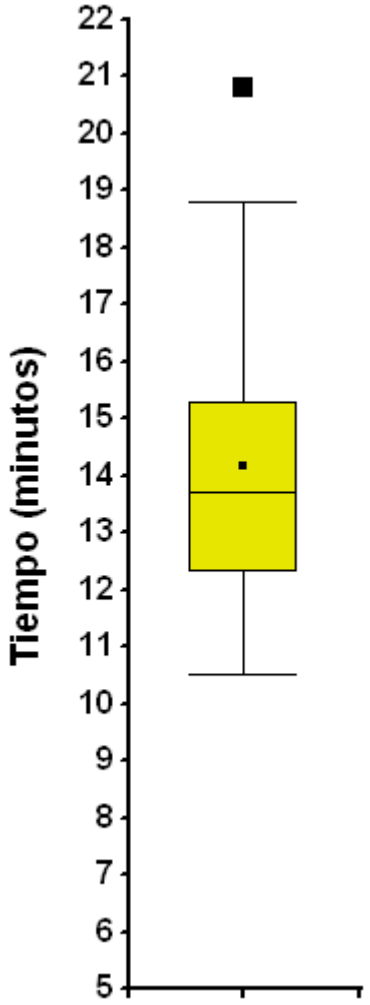
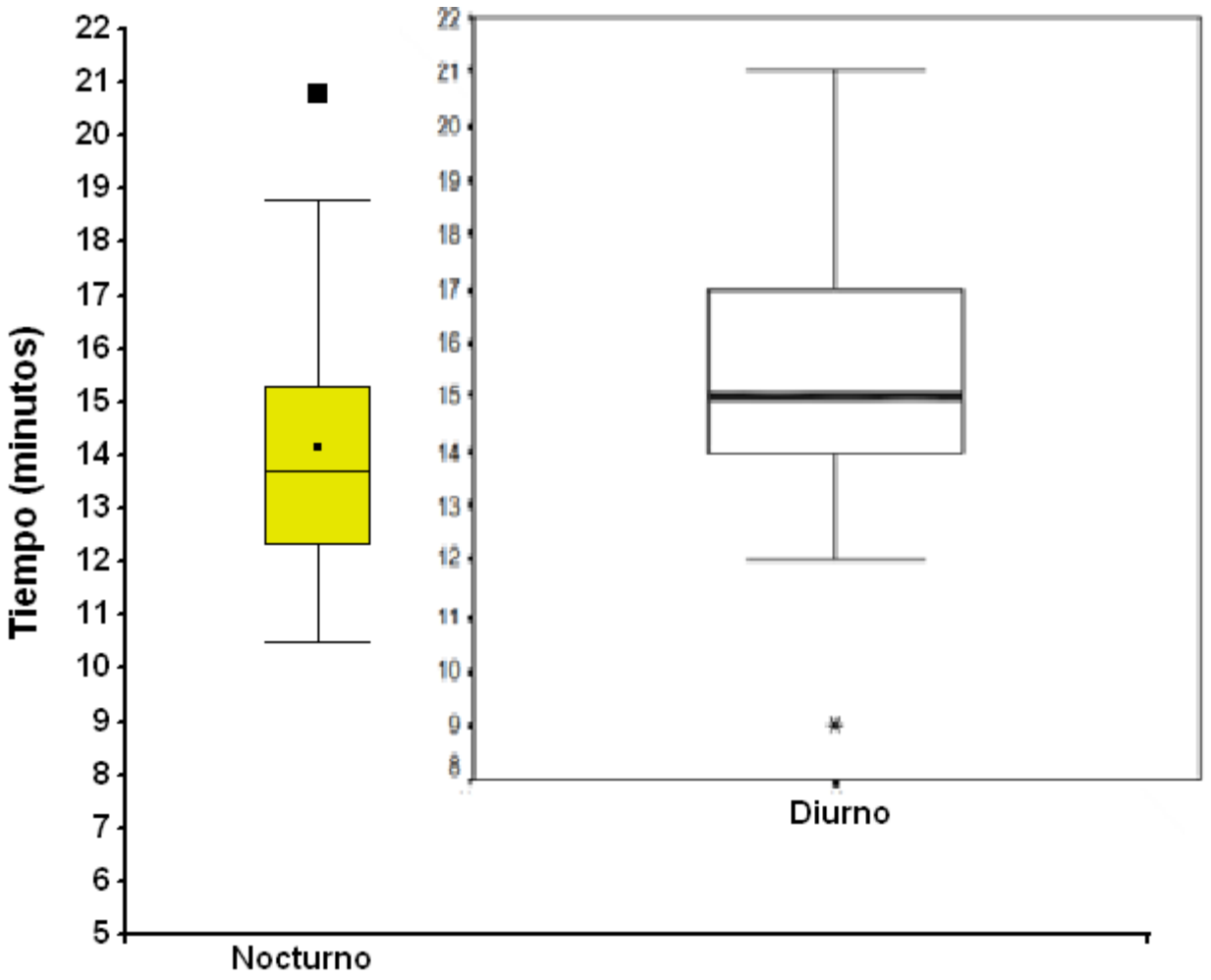
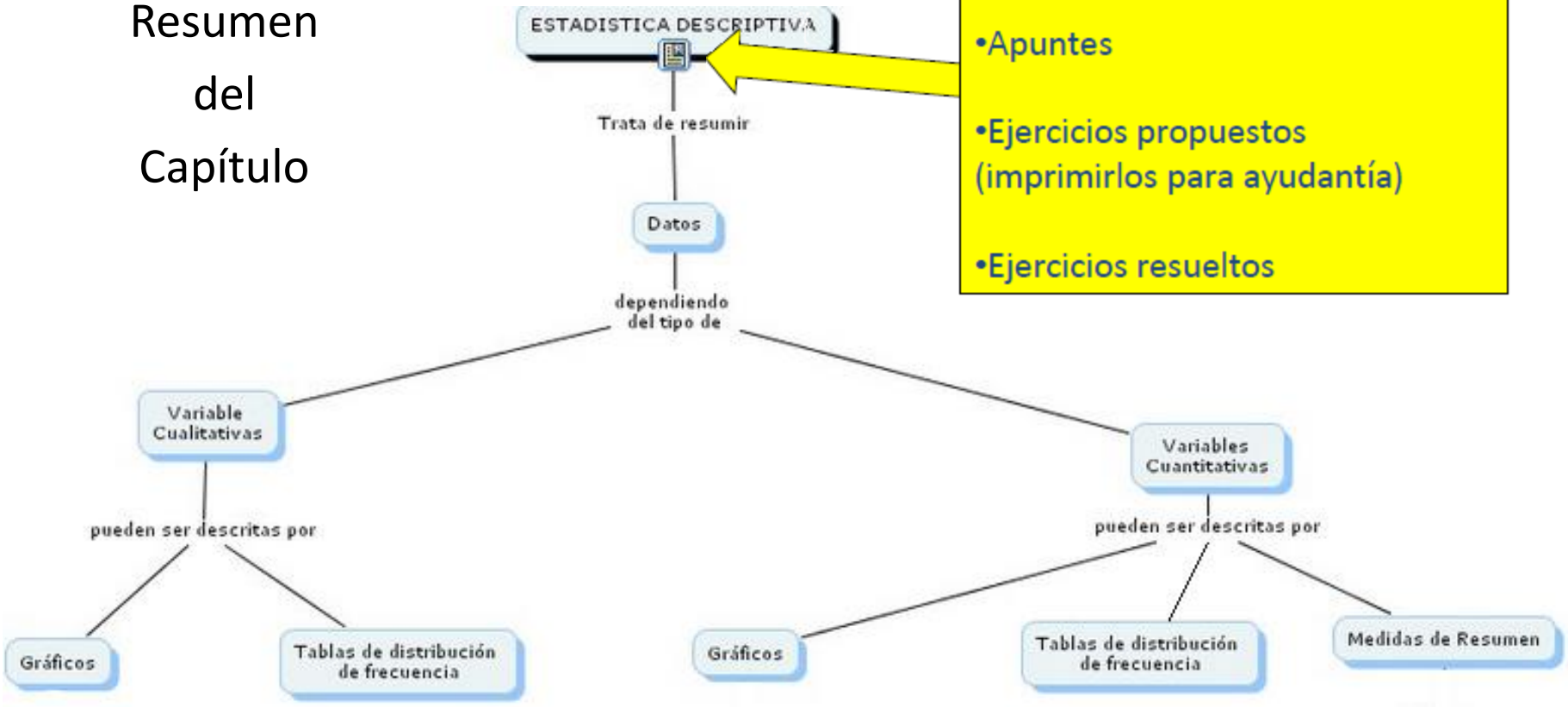


GRAFICO DE CAJA
Ejercicio 7 - Guía 2

A partir del Box Plot ¿Qué comparaciones es posible hacer entre los dos grupos?



Resumen del Capítulo



Aquí descarga:

- Apuntes
- Ejercicios propuestos (imprimirlos para ayudantía)
- Ejercicios resueltos

http://dta.utalca.cl/estadistica/

http://www.educandus.cl/estadistica/

Observación:

El gráfico de tallo y hoja no lo estudiamos, por lo tanto, tampoco los ejercicios que lo incluyan.