DESCRIPTIVA Y ANÁLISIS EXPLORATORIO DE DATOS

Estadística Descriptiva

- La Estadística Descriptiva, pretende dar una descripción numérica, ordenada y simplificada, a veces con la ayuda de representaciones gráficas, de la información obtenida en el relevamiento de datos de un fenómeno o situación en estudio.

- Población
- Muestra

- **Caracteres estadísticos: es una propiedad que permite clasificar a los individuos de una población. Se distinguen dos tipos:**
- ***a) Cualitativos***
- **Cualidades, no se pueden medir. Las modalidades son las diferentes situaciones de un carácter**
- ***b) Cuantitativos***
- **Son aquellos que se pueden medir o contar**

- **En el ordenamiento de los datos se deberá hacer la distinción entre datos (variables) de tipo continuo y discreto.**

- **La forma de la distribución de los datos (de una variable) se denomina *distribución de frecuencias*.**

- **En una población de N individuos, descrita según una variable o carácter X, cuyas modalidades han sido agrupadas vamos a definir:**
- ***Frecuencia absoluta :*** **Es el número de observaciones o sea es el número de veces que se repite dicho valor (f_i).**

- *Frecuencia absoluta acumulada :* **Es el número de elementos de la muestra cuya modalidad es inferior o equivalente al valor de la variable considerada (F_i).**
- *Frecuencia relativa :* **Es el cociente entre las frecuencias absolutas y el número total de observaciones o datos N**

$$h_i = \frac{f_i}{N}$$

- *Frecuencia relativa acumulada :* **Es el número de elementos de la muestra cuya modalidad es inferior o equivalente al valor de la variable considerada (F_i) dividido por el total de datos:**

$$H_i = \frac{F_i}{N}$$

- Como normalmente el conjunto de datos que se recolecta suele ser muy grande, es necesario disponer de alguna herramienta mediante la cual podamos visualizarlos.

- Para ello, una vez ordenados, hacemos un recuento de dichos datos y realizamos tablas estadísticas.

- En estas tablas, deberán figurar los valores de la variable en estudio, y sus frecuencias correspondientes.

- Si bien este ordenamiento puede evitarse al trabajar con programas específicos o alguno que posea este tipo de análisis, es útil para la realización de algunos gráficos.

- La principal dificultad para la obtención de una distribución de frecuencias, reside en la construcción de las modalidades, ya que ésta variará de acuerdo con el tipo de variable que se pretende describir: si la variable es cualitativa, se tomarán como modalidades las distintas respuestas observadas de la muestra.

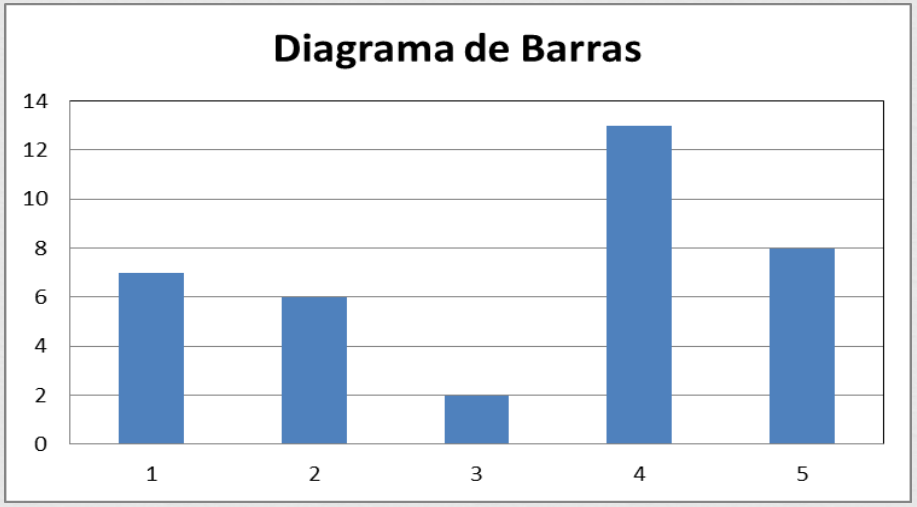
- Si la variable es discreta (que tome pocos valores distintos), las modalidades coincidirán con los distintos valores medidos en la muestra.
- Si la variable es continua (o bien discreta, pero toma muchos valores distintos), se tomarán como modalidades intervalos de clase. Los intervalos donde se encuentran los datos agrupados, se simbolizan por $[L_{i-1}, L_i)$.

- Luego dependiendo del tipo de agrupamiento o del tipo de dato que se tenga las definiciones de los distintos tipos de frecuencia son equivalentes.

- Cuando en una serie de datos no se repiten valores es decir no hay frecuencias, la serie será tratada como serie simple.

Gráficos

- Gráficos para variables cualitativas o atributos
- Diagrama de barras o bastones



- **Diagramas de sectores:**

- Se utilizan para hacer comparaciones de las distintas modalidades de un carácter mediante sectores circulares

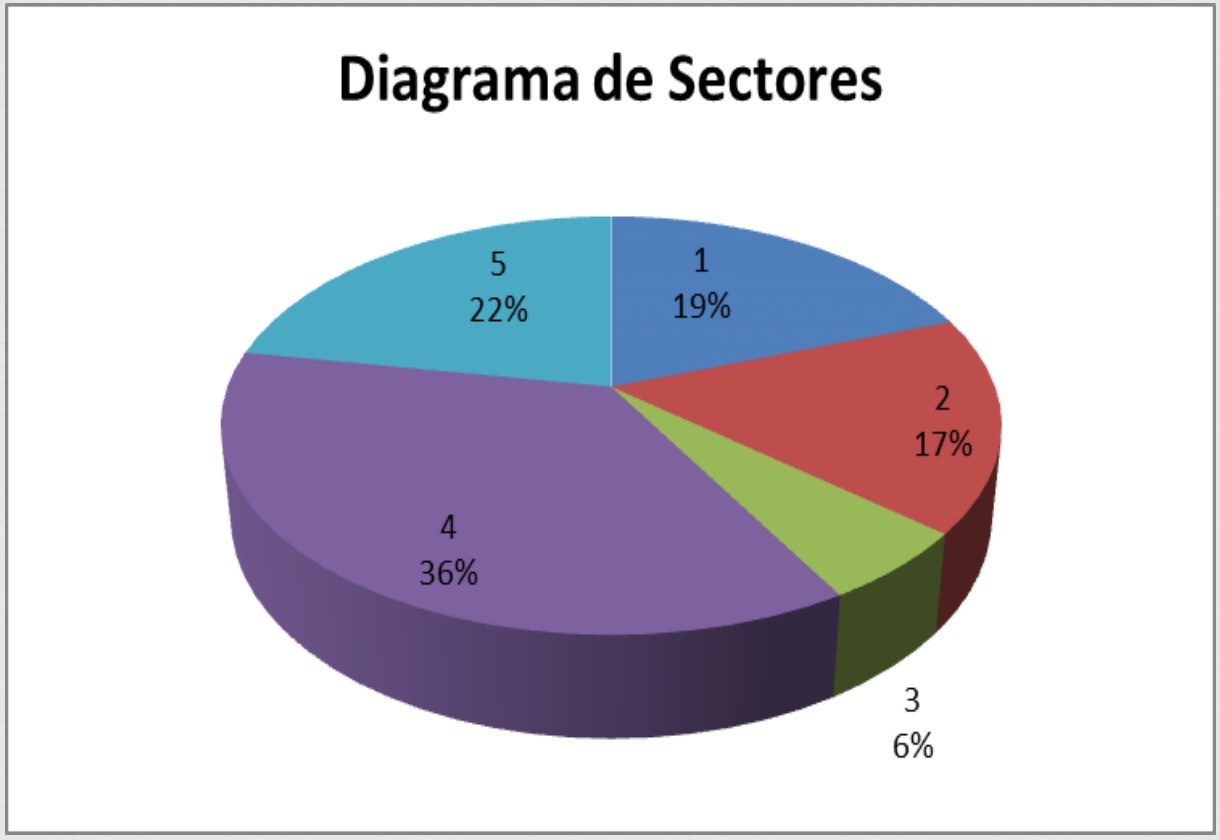
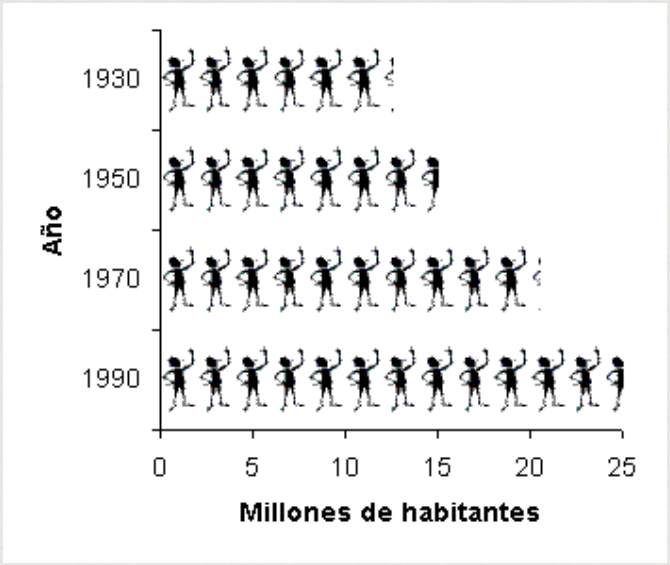


Diagrama de Sectores
5
22%
1
19%
2
17%
4
36%
3
6%

- **Pictogramas:** el perímetro del dibujo tiene que ser proporcional a la frecuencia, pero esto puede llevar a un efecto visual engañoso ya que a frecuencia doble corresponde un dibujo de área cuádruple, con lo cual tiene un inconveniente debido a la falta de precisión.


Árboles plantados
150
100
50
0
Enero
Febrero
Marzo
Abril
Mayo
Junio
Julio
Agosto
Septiembre
Octubre
Noviembre
Diciembre

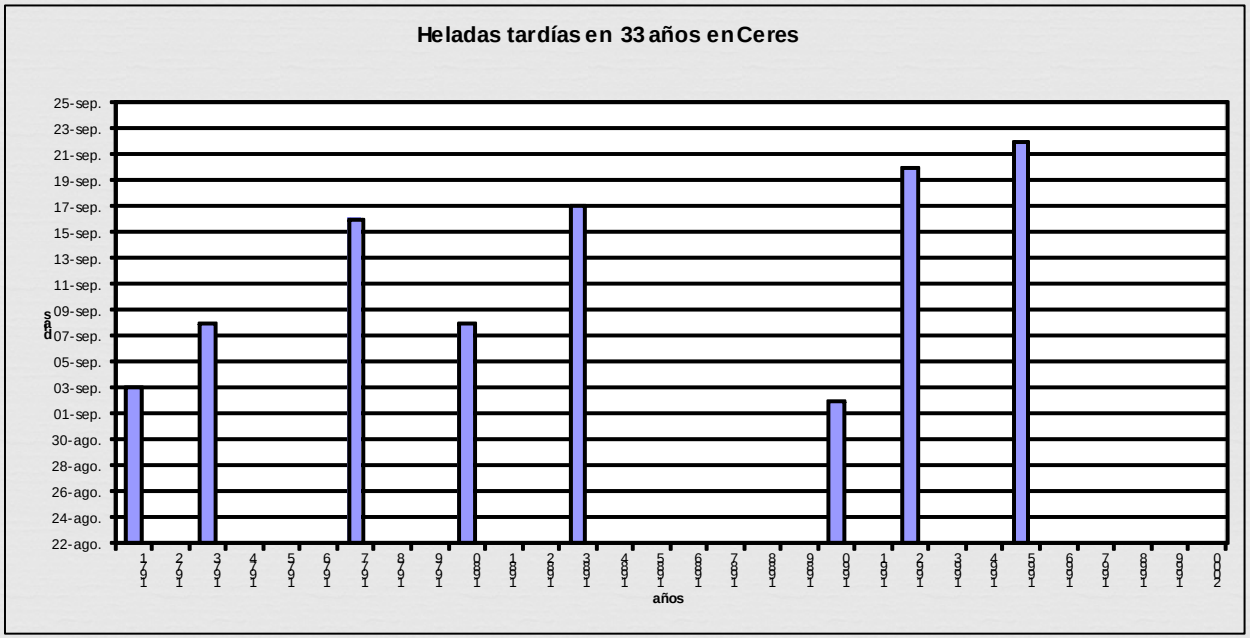

Año
1930
1950
1970
1990
0
5
10
15
20
25
Millones de habitantes

Gráficos para Variables Cuantitativas

- Variables discretas
  - **Diagrama de barras**

- Variables continuas
  - **Histograma**

Diagrama de barras

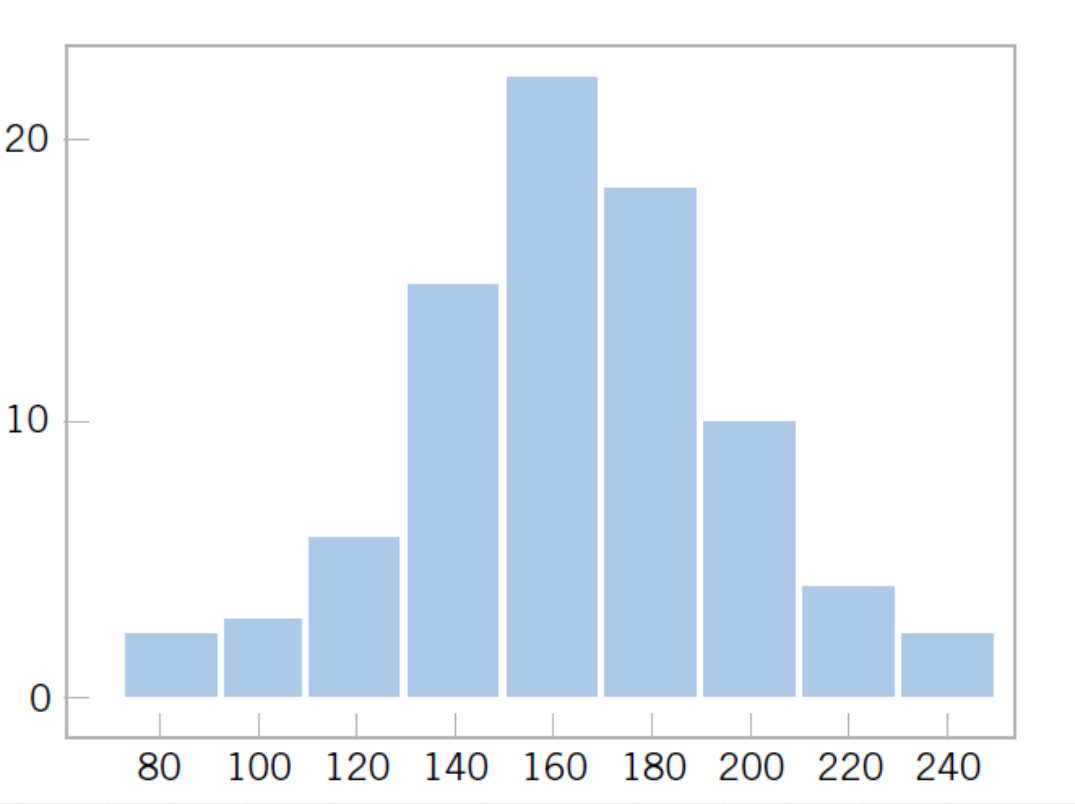


Histograma

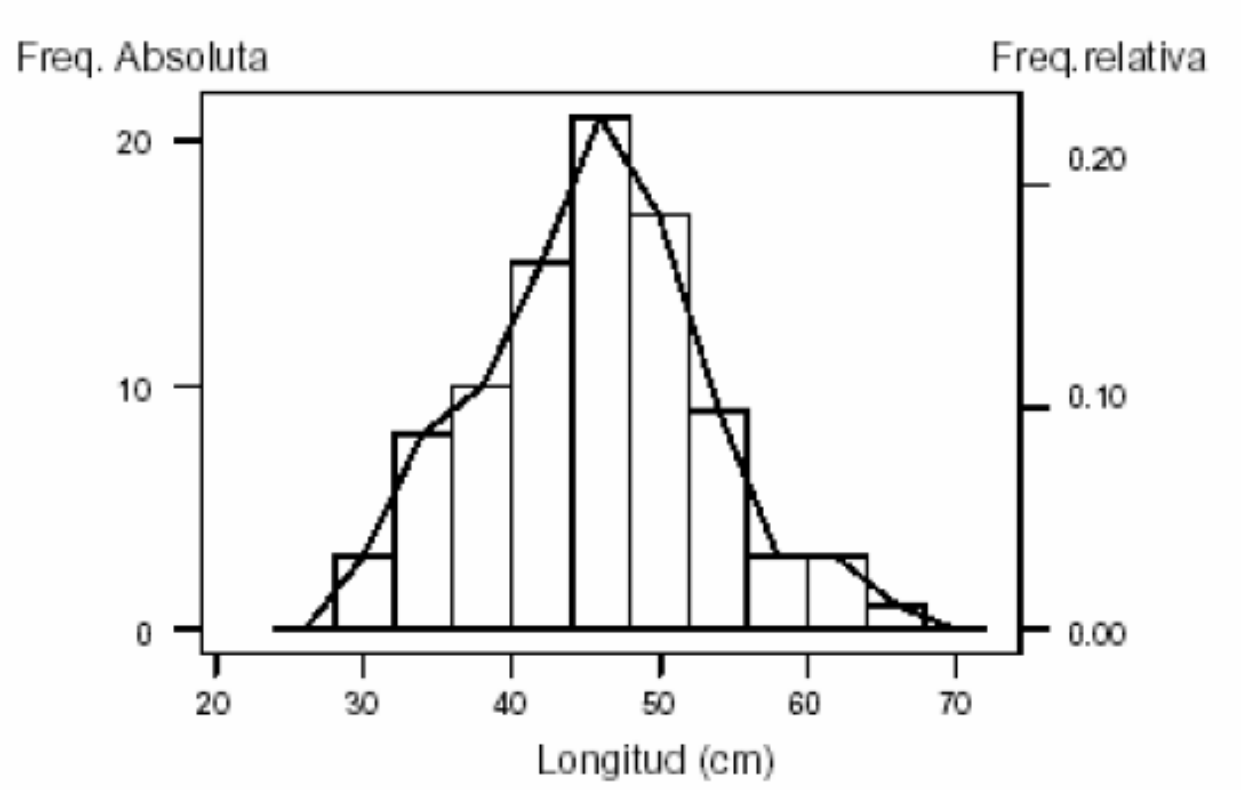
- 1.- *¿Cuántos intervalos construir?*
- 2.- *¿Qué valor se elige como extremo inferior del primer intervalo* L_0 *?*

Consejos:

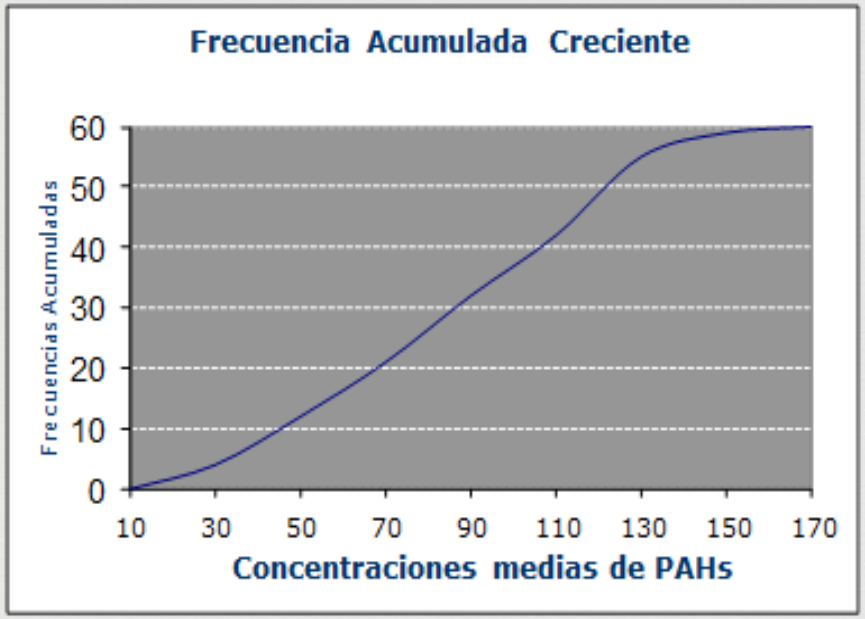
- 1. Usar intervalos de la misma longitud
- 2. Los intervalos no pueden solaparse
- 3. Cada observación sólo puede pertenecer a un intervalo
- 4. Todos los datos deben pertenecer a algún intervalo

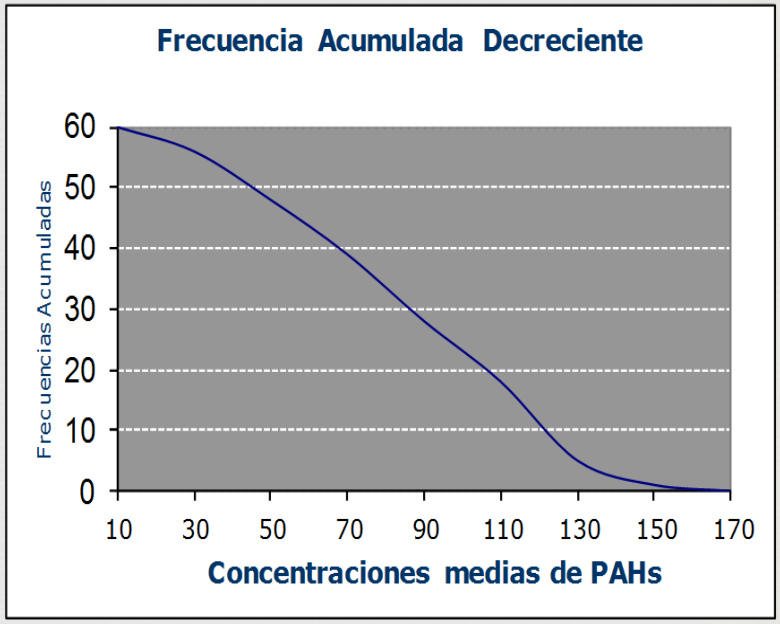

20
10
0
80
100
120
140
160
180
200
220
240

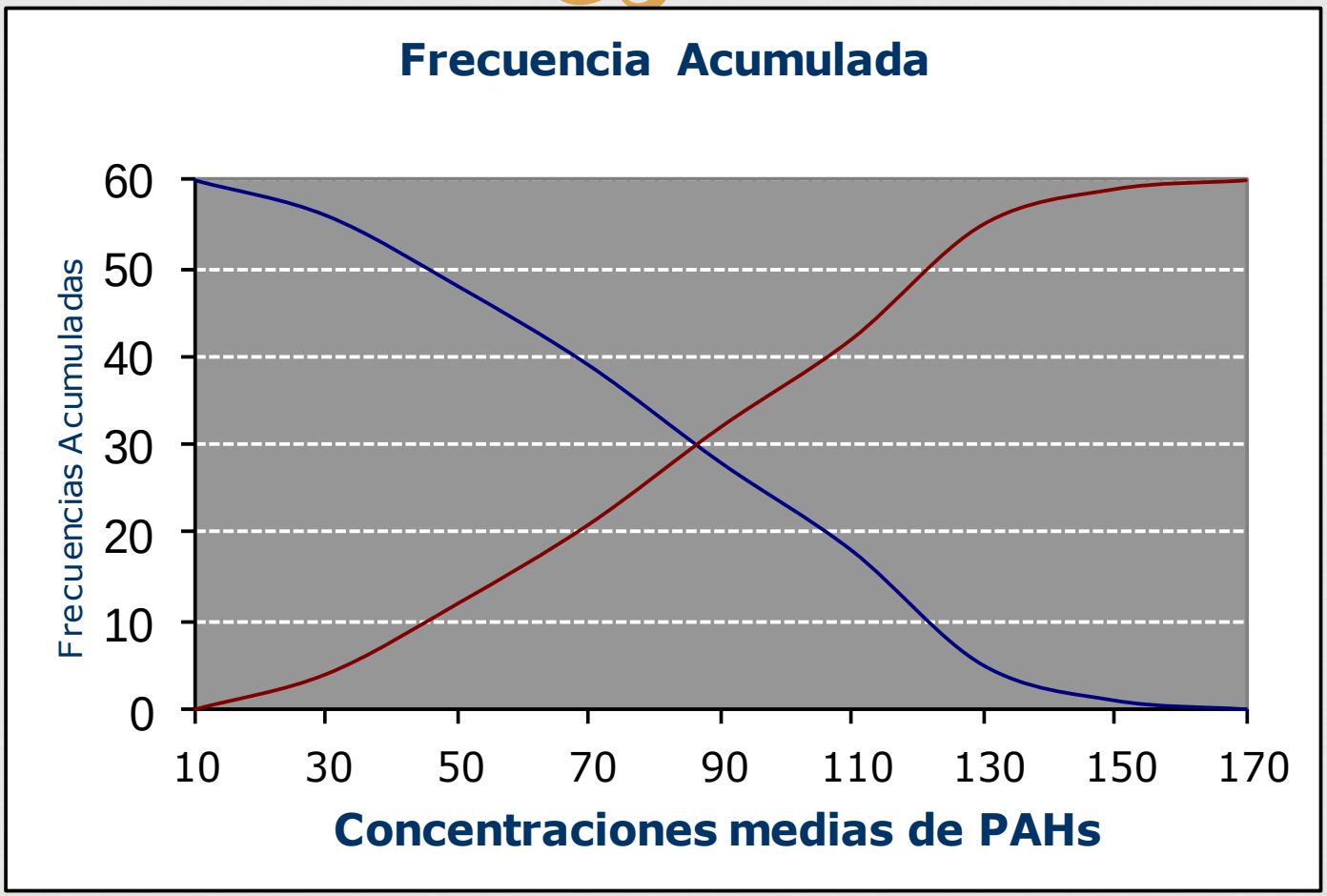
Polígono de frecuencias



Gráficos de Frecuencias acumuladas






Frecuencia Acumulada
Frecuencias Acumuladas
60
50
40
30
20
10
0
10
30
50
70
90
110
130
150
170
Concentraciones medias de PAHs

Análisis Exploratorio

- Análisis reciente, son métodos innovadores para el análisis de datos.
- Hace énfasis en la exploración de los datos por métodos gráficos previos al clásico análisis estadístico.

- La visualización de los datos permite al investigador penetrar en su estructura, minimizando los supuestos probabilísticos que tradicionalmente se asumen con respecto a su comportamiento y distribución. Lo anterior equivale a proporcionarle al investigador "una lente" de aumento que le permite:

- Exhibir características o patrones ocultos dentro de los datos.
- Resaltar con claridad la tendencia que conforman los datos.
- Proporcionar hipótesis o modelos acerca del comportamiento de los datos
- Se ha robustecido con la reciente aparición de diversos programas como por ejemplo Statgraphics, Statistica, SPLUS, etc .

Herramientas más importantes :

- El diagrama de tallo y hoja.
- -El diagrama de caja.
- Las profundidades.
- El diagrama de letras.
- Transformaciones matemáticas.
- Suavizaciones.
- Análisis de series de tiempo.

El A.E.D. proporciona:

- métodos sistemáticos sencillos para organizar y preparar los datos
- detectar fallos en el diseño y recogida de los mismos, tratamiento y evaluación de datos ausentes (missing),
- identificación de casos atípicos (outliers) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad).

- El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida.

- **Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.**

El Diagrama de Tallo y Hoja

- Combina los aspectos visuales del histograma con la información numérica que proporciona una tabla de distribución de frecuencias.

- Es un gráfico muy sencillo de realizar, se puede considerar como la técnica de representación gráfica recomendable para variables cuantitativas, por encima de otra forma muy usual como el histograma.

Construcción

- 1.-Ordenar el lote de datos en magnitud creciente.
- 2. Fraccionar en dos partes el dato según la característica de los datos o lo que se quiere mostrar de ellos.
- 3. Formar el tallo (parte más significativa del número) y las hojas (el resto de las cifras) con las fracciones respectivas.
- 4. Construir el tallo escribiendo verticalmente los dígitos enteros ordenados en forma creciente, asociando a cada uno su hoja respectiva.

- En términos generales hace visibles las siguientes características:
- 1. Muestra el rango de valores que los datos cubren.
- 2. Determina donde se concentran la mayoría de los datos
- 3. Describen la simetría del conjunto de datos.
- 4. Identifica si existen huecos en la distribución de los datos.
- 5. Señala aquellos valores que claramente se desvían del conjunto de datos.

- La observación de cualquiera de estos gráficos: el histograma o el diagrama de tallo y hoja, permite extraer ideas de las características generales de la variable representada.

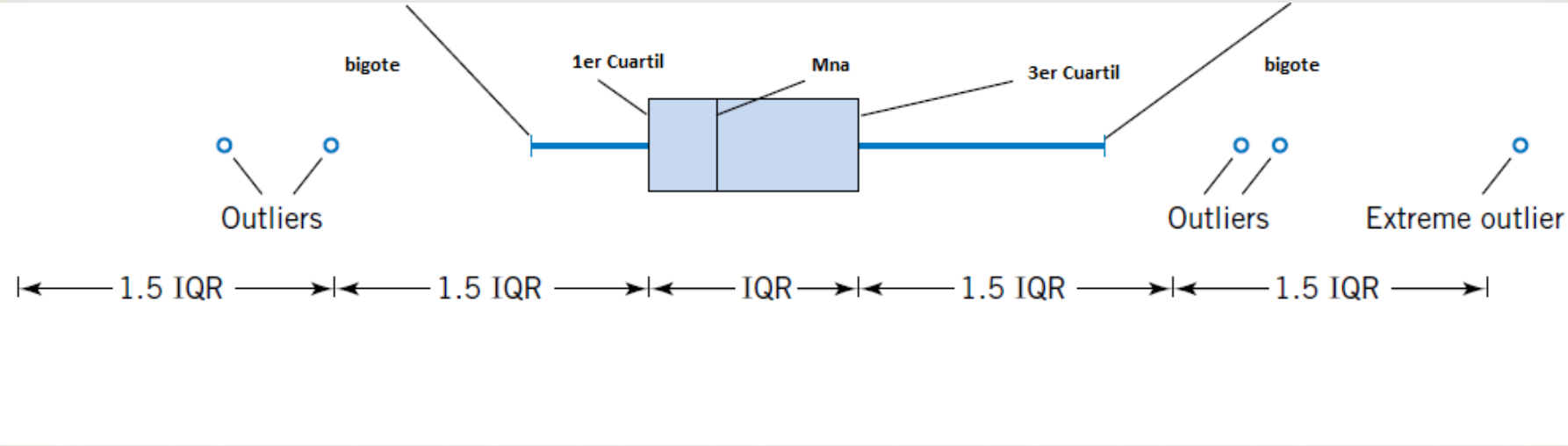
0 | 99
1 | 0011112222233333333444444
1 | 556778
2 | 011222334
2 | 677888899
3 | 00111122234
3 | 5568899
4 | 011122224444
4 | 55566677788888
5 | 0012

1	0|5
7	0|666777
26	0|88899999999999999
(20)	1|00000000001111111111
43	1|2233333
36	1|44444444444555555
19	1|6666666677
9	1|889
6	2|000
3	2|22

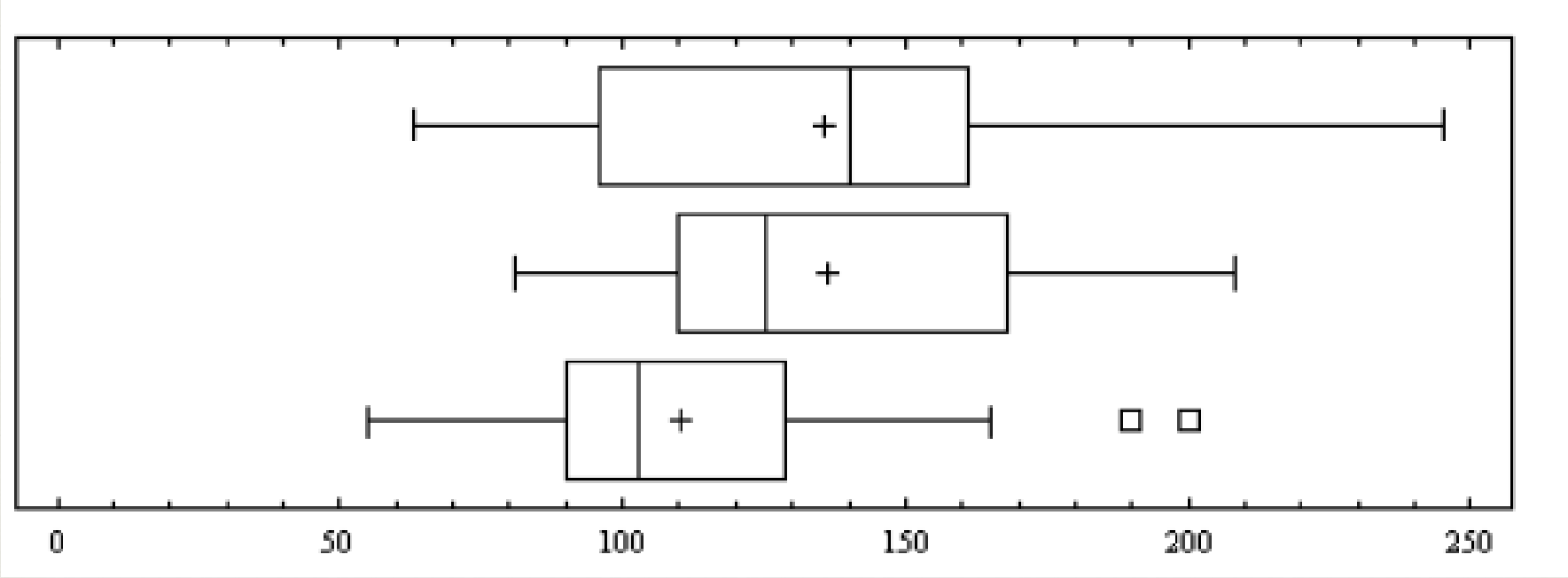
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

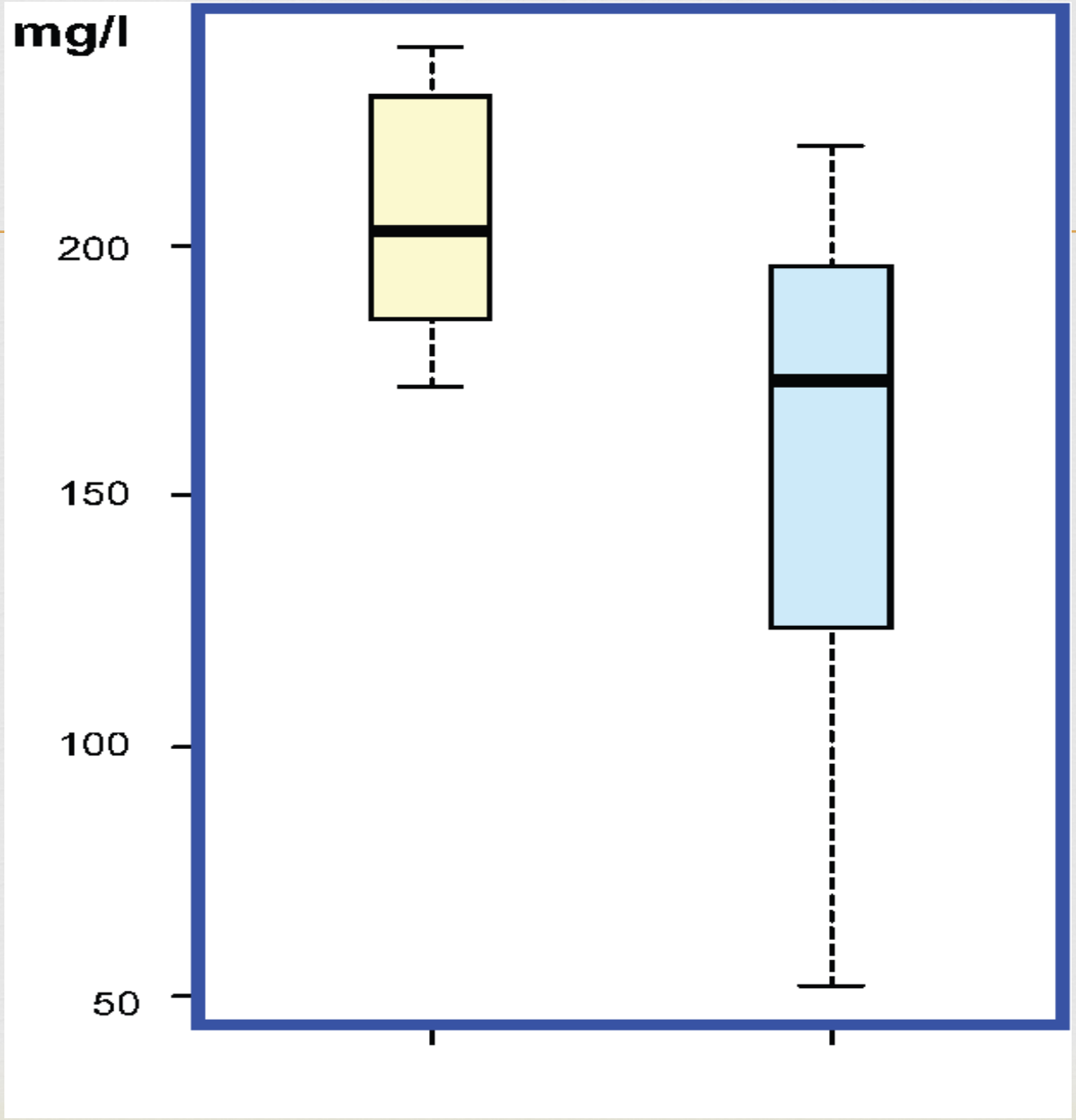
Gráfico de caja y bigote

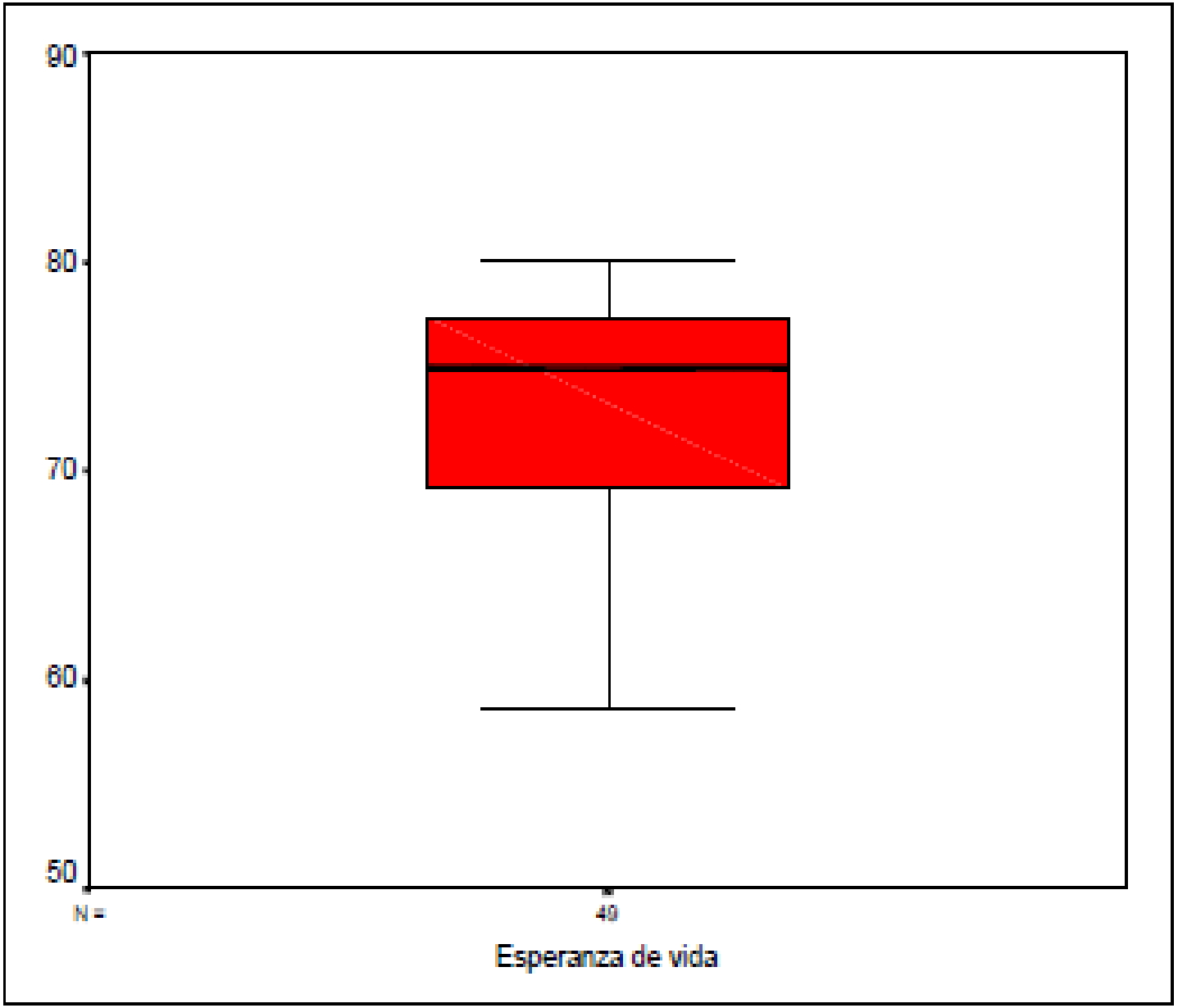
- Es un gráfico basado en cinco datos para construirlo: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, y el valor máximo. Ayuda a visualizar un conjunto de datos.


bigote
1er Cuartil
Mna
3er Cuartil
bigote
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR


mediana
mín
Q1
Q3
máx
12
14
16
18
20
22
24
26
28
30
32


0
50
100
150
200
250


mg/l
200
150
100
50


90
80
70
60
50
N =
49
Esperanza de vida

- Es posible introducir algunas variaciones en la construcción de estos diagramas, dependiendo del tipo de estudio y de la información disponible.
- La caja o rectángulo contiene un porcentaje de la muestra y puede construirse con diferentes rangos de variación.
- Es recomendable señalar con una marca los valores atípicos.

CARACTERÍSTICAS de una muestra

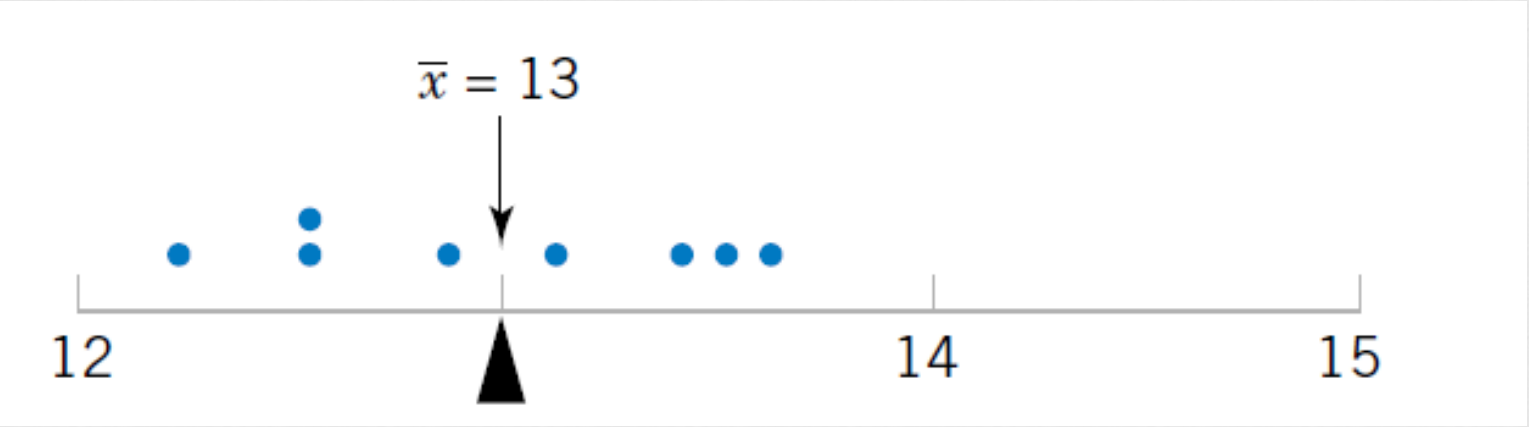
- MEDIDAS DE TENDENCIA CENTRAL
- MEDIDAS DE DISPERSIÓN
- MEDIDAS DE FORMA
  - *ASIMETRIA*
  - *CURTOSIS*

MEDIDAS DE TENDENCIA CENTRAL

Promedios

- Media aritmética o media de muestra:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$


$\bar{x} = 13$
12
14
15

Propiedades

$$n.\bar{x} = \sum x_i$$

$$\sum (x_i - \bar{x}) = 0$$

$$\sum (x_i - \bar{x})^2 = mínimo$$

$$\bar{X} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \bar{x}_n N_n}{N}$$

Media Geométrica

$$Gm = \sqrt[n]{\Pi\ x_i}$$

$$\log G_m = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)$$

Propiedades

- está menos afectada por valores extremos.
- para cualquier serie es siempre menor que la media aritmética.
- es muy útil en el cálculo de *números índice*.
- se puede manipular algebraicamente.
- no es muy conocida y no puede evaluarse cuando hay datos negativos o ceros.

Media armónica

$$\frac{1}{Hm} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{N}$$

- El empleo de la media geométrica o de la armónica equivale a una transformación de la variable en log x ó 1/x , respectivamente, y el cálculo de la media aritmética de la nueva variable; por ejemplo, si la variable abarca un campo de variación muy grande, tal como el porcentaje de impureza de un producto químico, por lo general alrededor del 0.1%, pero que en ocasiones llega incluso al 1% o más, puede ser ventajoso el empleo de log x en lugar de x para obtener una distribución más simétrica.

$$H \leq G \leq \overline{X}$$

Medidas de ubicación

- **Modo: es el valor que se corresponde con la máxima frecuencia.**
- **Si hay un gráfico de intervalos se busca interpolar:**

$$Mo = L_{iMo} + \frac{d_1}{d_1 + d_2} c$$

$$Mo = L_{iMo} + \frac{f_1}{f_1 + f_2} c$$

Mediana

- **Variables discretas:**

- Si no hay frecuencias

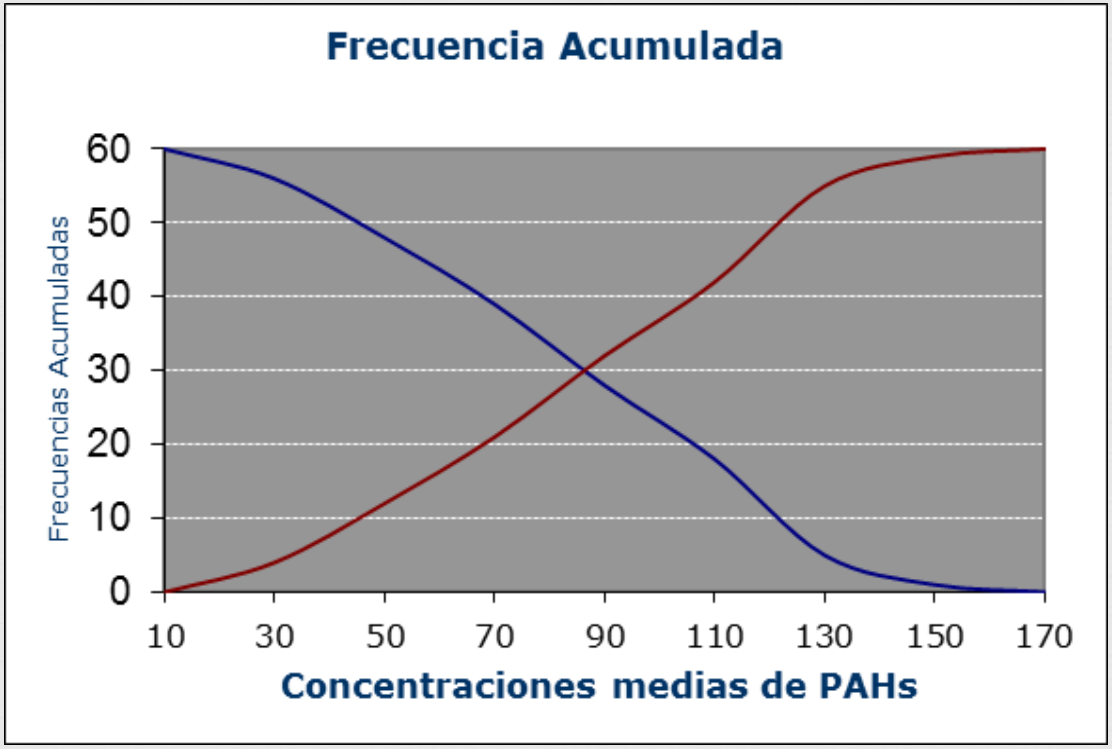
- - Si el número de datos es impar la Mna. es el valor central.
- - Si el número de datos es par laMna. es la semisuma de los valores centrales.

- Si hay frecuencias:

- - Se calcula N/2 y obtienen las frecuencias acumuladas (Ni)
- - Se observa cual es la primera Ni que supera o iguala a N/2, distinguiéndose dos casos:
- - Si existe un valor de x_i tal que $N_{i-1} < N/2 < N_i$, entonces se toma como $Mna. = x_i$
- - Si existe un valor i tal que $N_i = N/2$ entonces la mediana será
- $$Mna. = \frac{x_i + x_{i+1}}{2}$$

- Variables continuas: se obtiene interpolando

$$Mediana = L_i + \frac{N/2 - FL_i}{f_i}\, c$$


Frecuencia Acumulada
Frecuencias Acumuladas
60
50
40
30
20
10
0
10
30
50
70
90
110
130
150
170
Concentraciones medias de PAHs

Propiedades

- No está influenciada por valores extremos. Por lo tanto, es una medida conveniente de la ubicación central.

- -Un valor seleccionado a azar se ubicará por arriba o por debajo de ella con igual probabilidad; por esto suele llamársela valor probable.

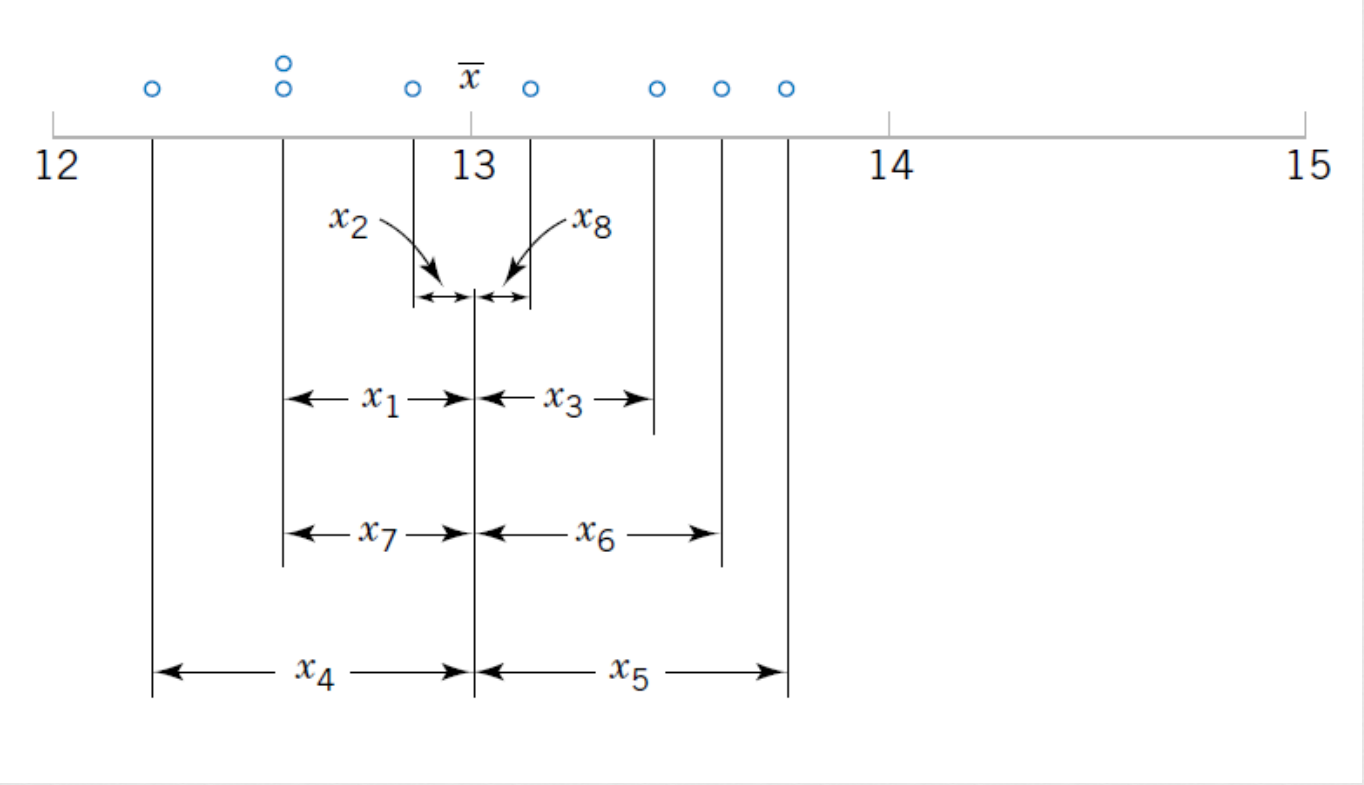
- Algunas desventajas son:

- -No se la puede manipular algebraicamente.
- -No es tan usada como la media aritmética, y tiene mayor error que ella.

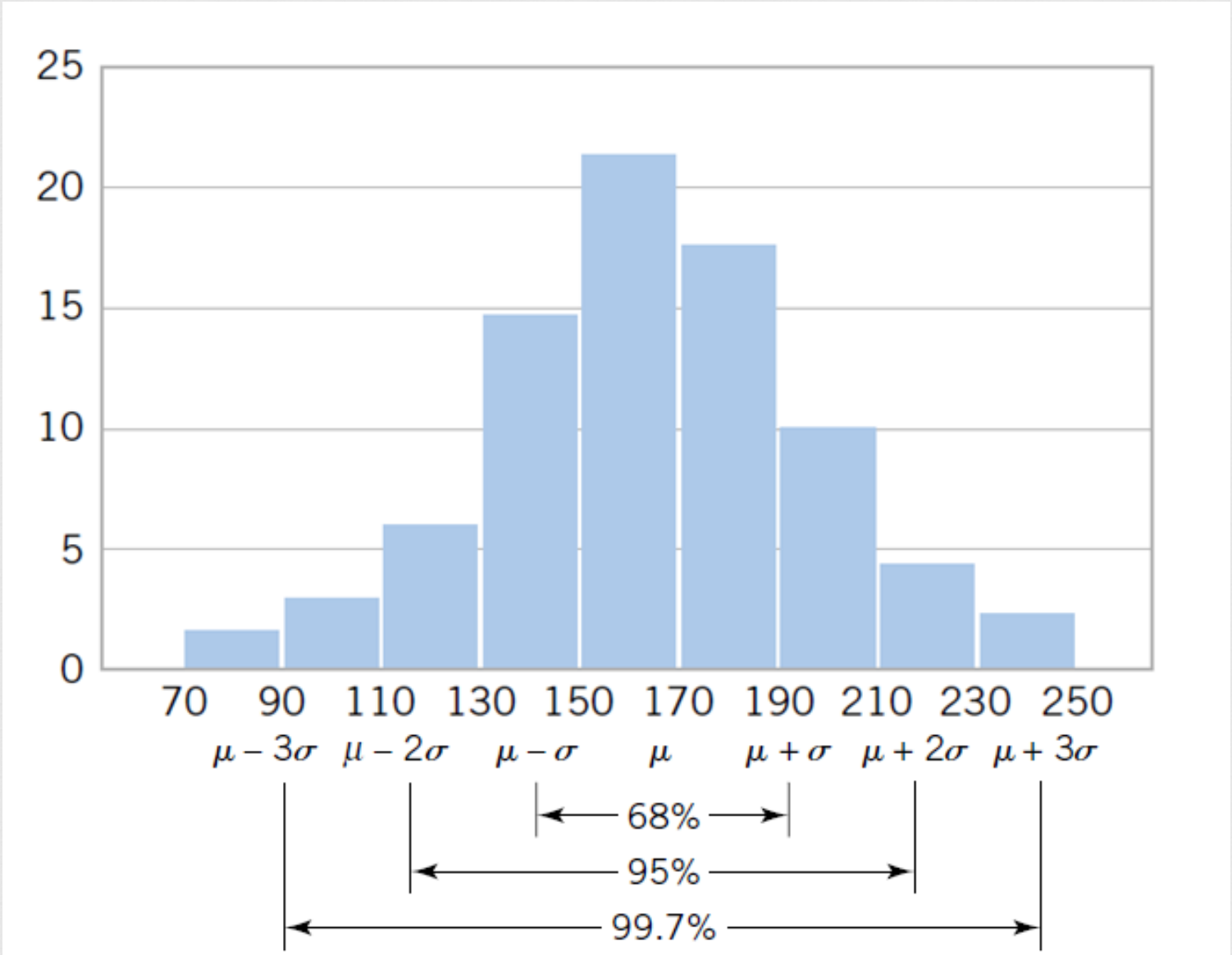
MEDIDAS DE DISPERSION

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$


$\overline{x}$
12
13
14
15
x_2
x_8
x_1
x_3
x_7
x_6
x_4
x_5

- Se utiliza una regla empírica para interpretar los valores de la varianza o desvío, se usará cuando la muestra sea grande y la forma de la muestra sea aproximadamente de campana.
- Esta regla considera que si se miden en el eje x y hacia ambos lados de la media una distancia igual al desvío, en ese intervalo quedarán comprendidos el 68% de las observaciones.
- Si se traza dos veces el desvío hacia ambos lados de la media quedarán comprendidos el 95% de las observaciones en ese intervalo.
- Si se trazan tres veces el desvío quedarán comprendidos el 99% de las observaciones entre esos límites.


25
20
15
10
5
0
70
90
110
130
150
170
190
210
230
250
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
68%
95%
99.7%

MEDIDAS DE FORMA

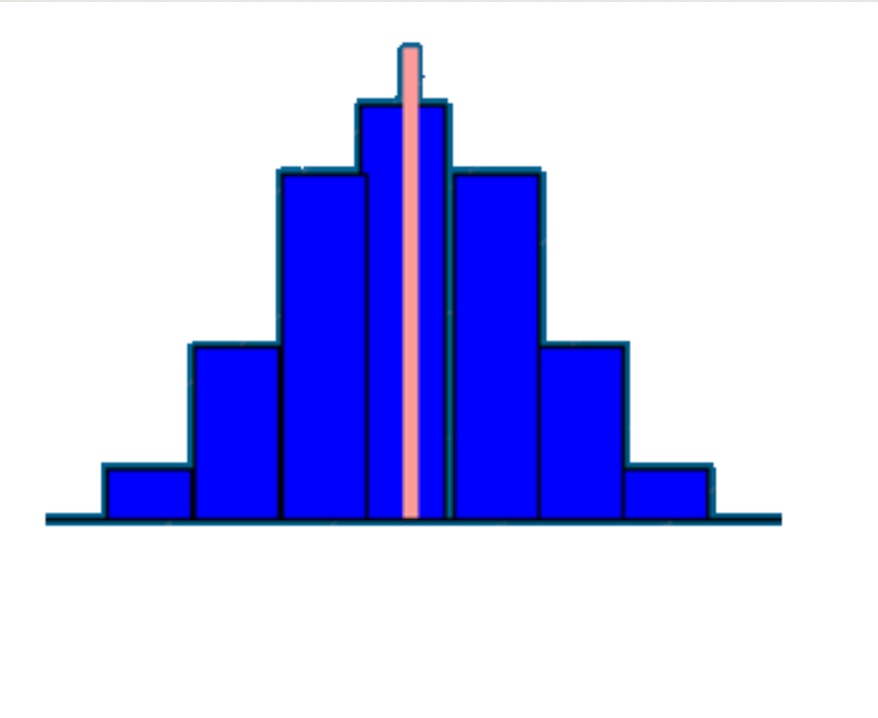
ASIMETRÍA

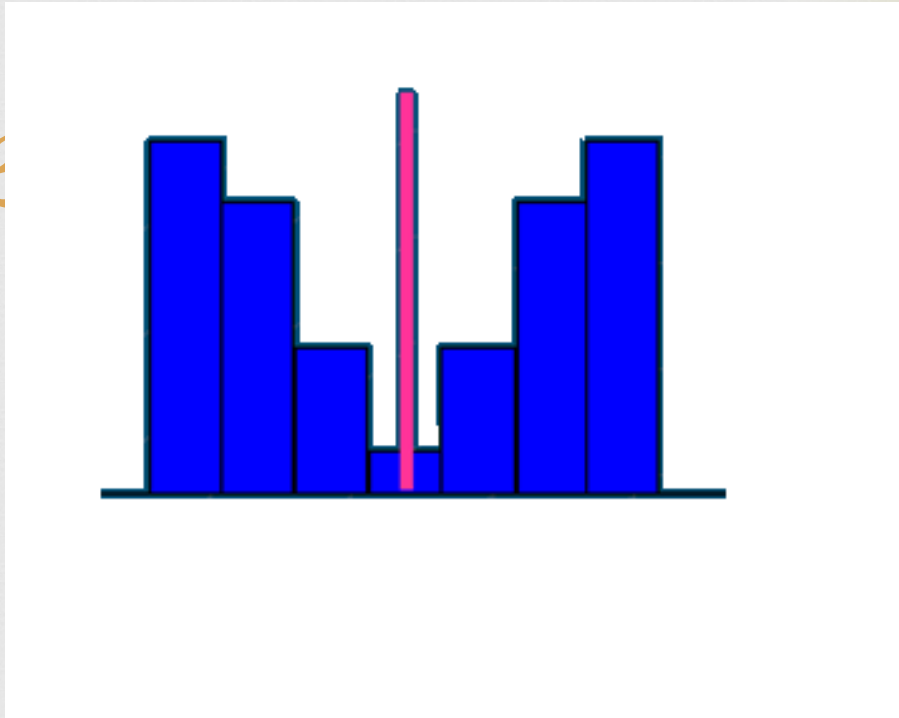
$$As = \frac{(\bar{x} - Modo)}{S}$$

$$As = \frac{3(\bar{x} - Mediana)}{S}$$

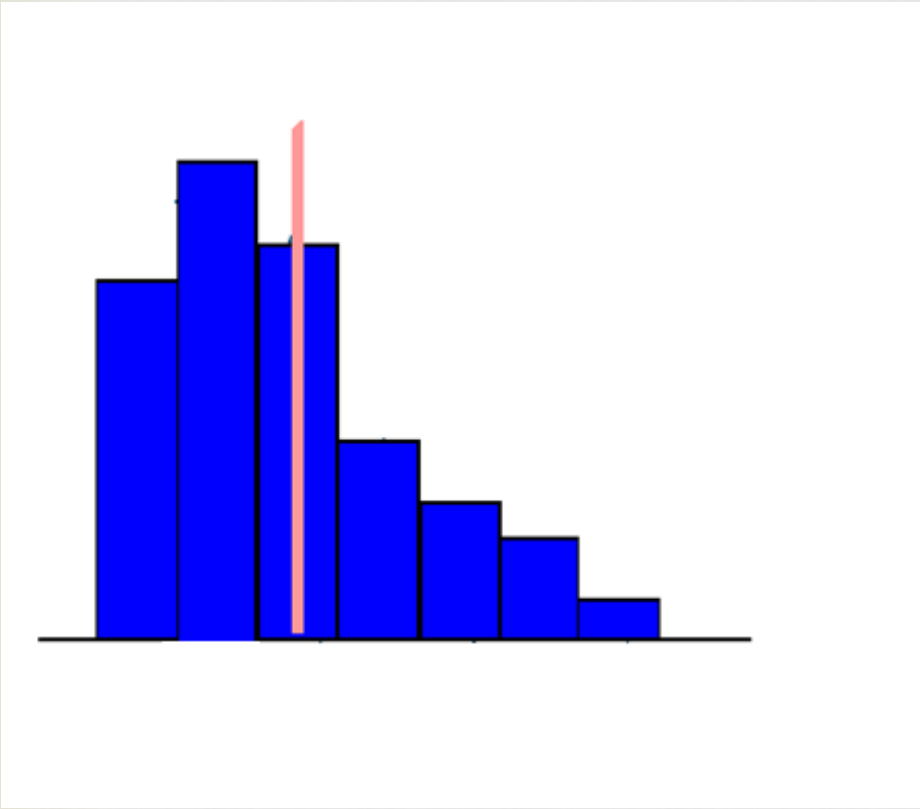
$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n}$$

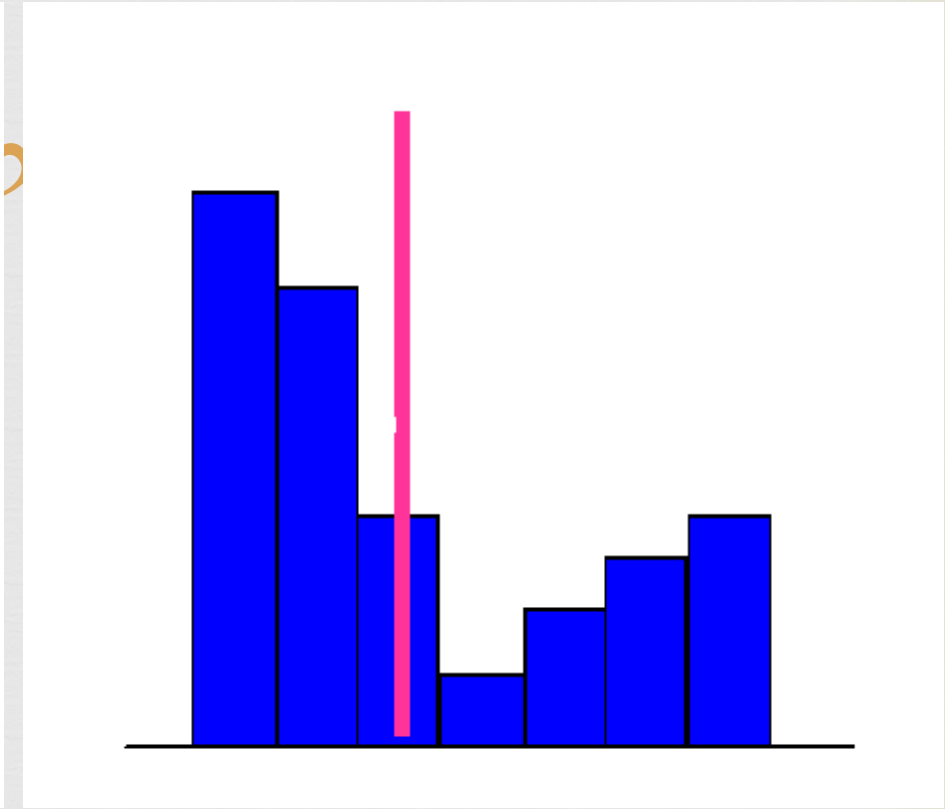
$$As = \frac{m_3}{S^3}$$



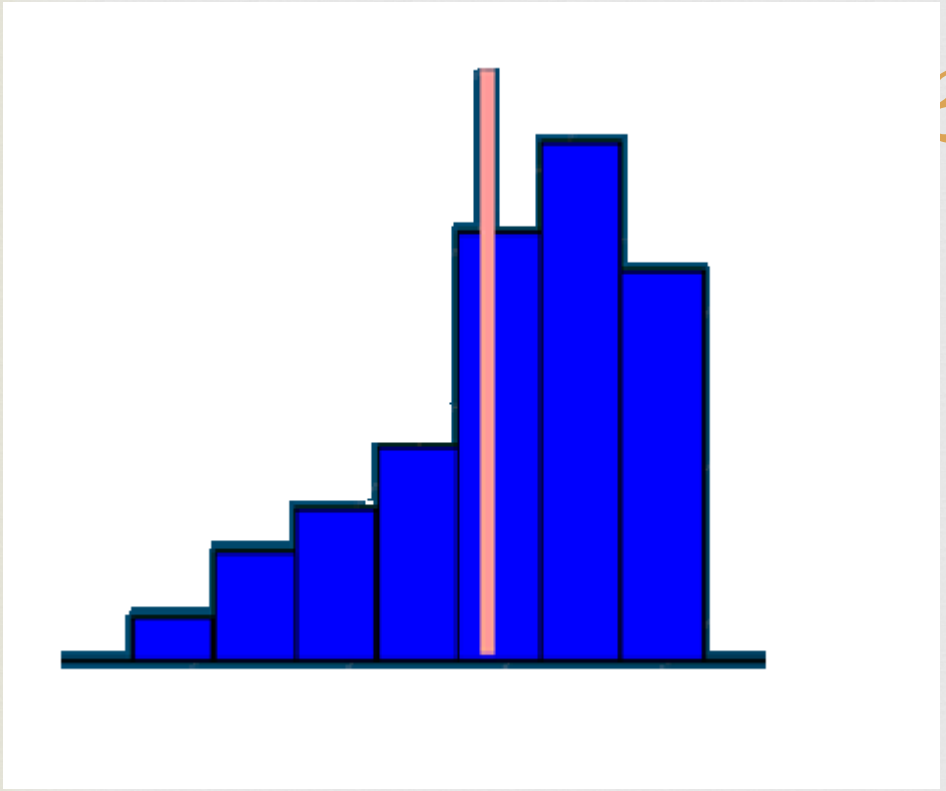


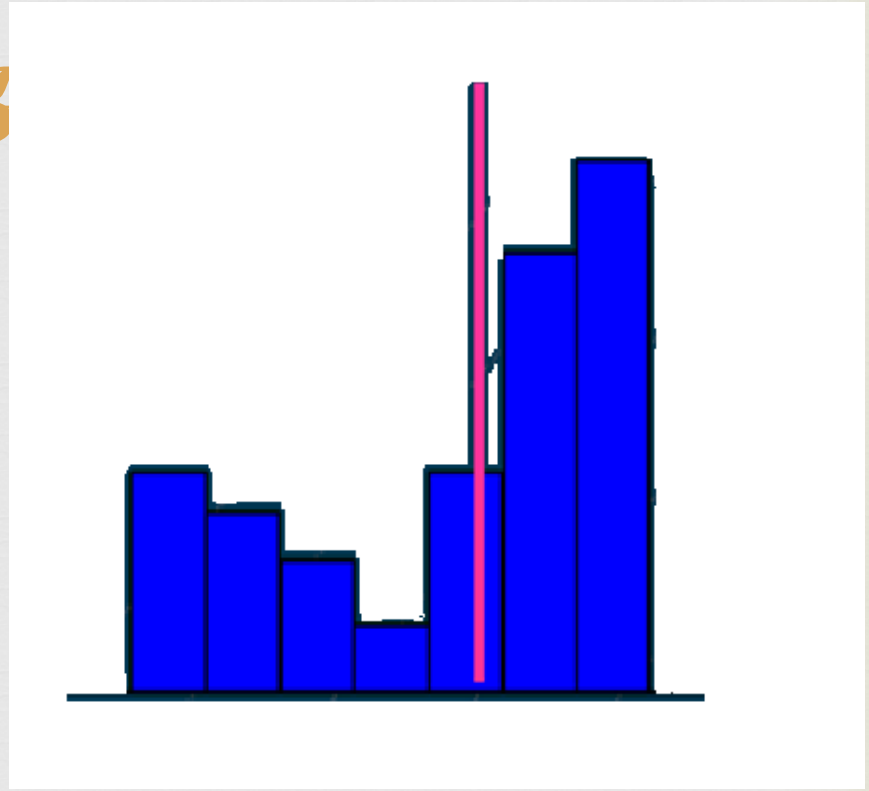
Distribución simétrica





Distribución asimétrica positiva o a la derecha





Distribución asimétrica negativa o a la izquierda

CURTOSIS

$$K = \frac{1}{2}\frac{(Q_3 - Q_1)}{(P_{90} - P_{10})}$$

$$K = \frac{m_4}{S^4}$$

Medidas Descriptivas Numéricas y Representaciones Graficas aconsejadas en función de la escala de medida de la variable

Escala de medida	Representaciones gráficas	Medidas de tendencia central	Medidas de dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación Típica
Razón		Media Geométrica	Coeficiente de Variación