



1970  
2020

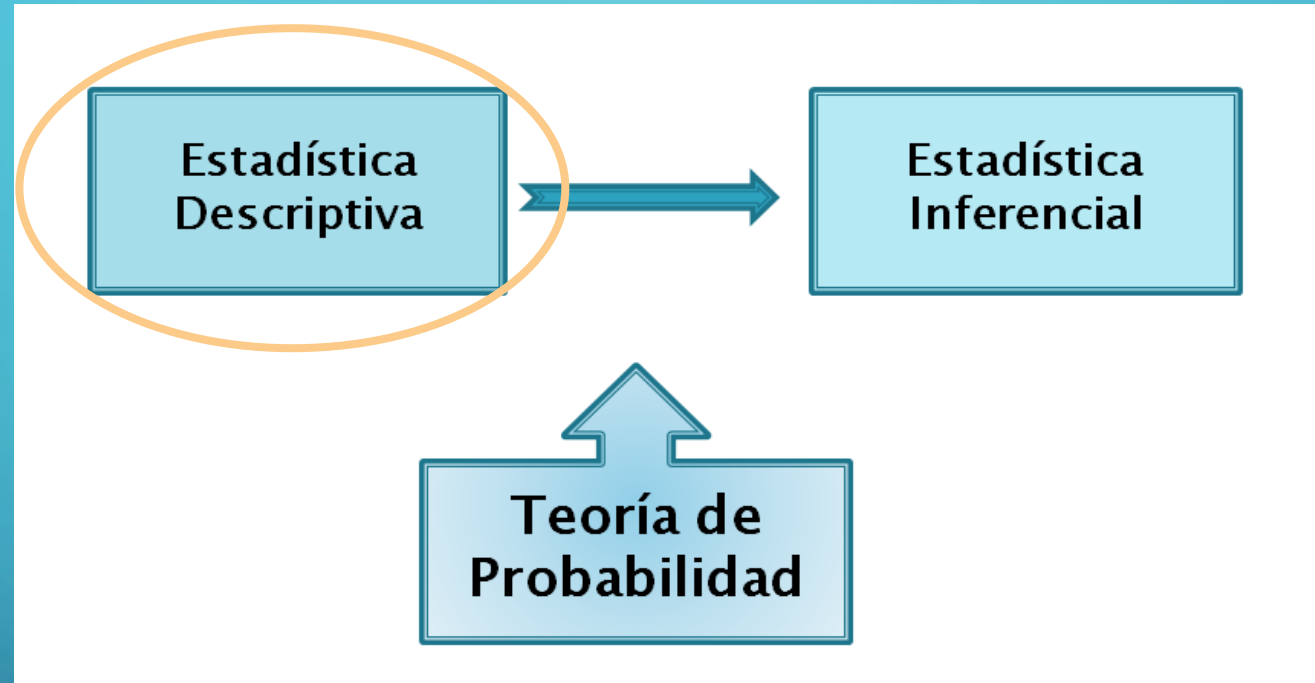
FICH

- UNIVERSIDAD NACIONAL DEL LITORAL
- FACULTAD DE INGENIERÍA Y CIENCIAS  
HÍDRICAS

# • ESTADÍSTICA

- INGENIERÍA EN INFORMÁTICA

• *MG. SUSANA VANLESBERG*



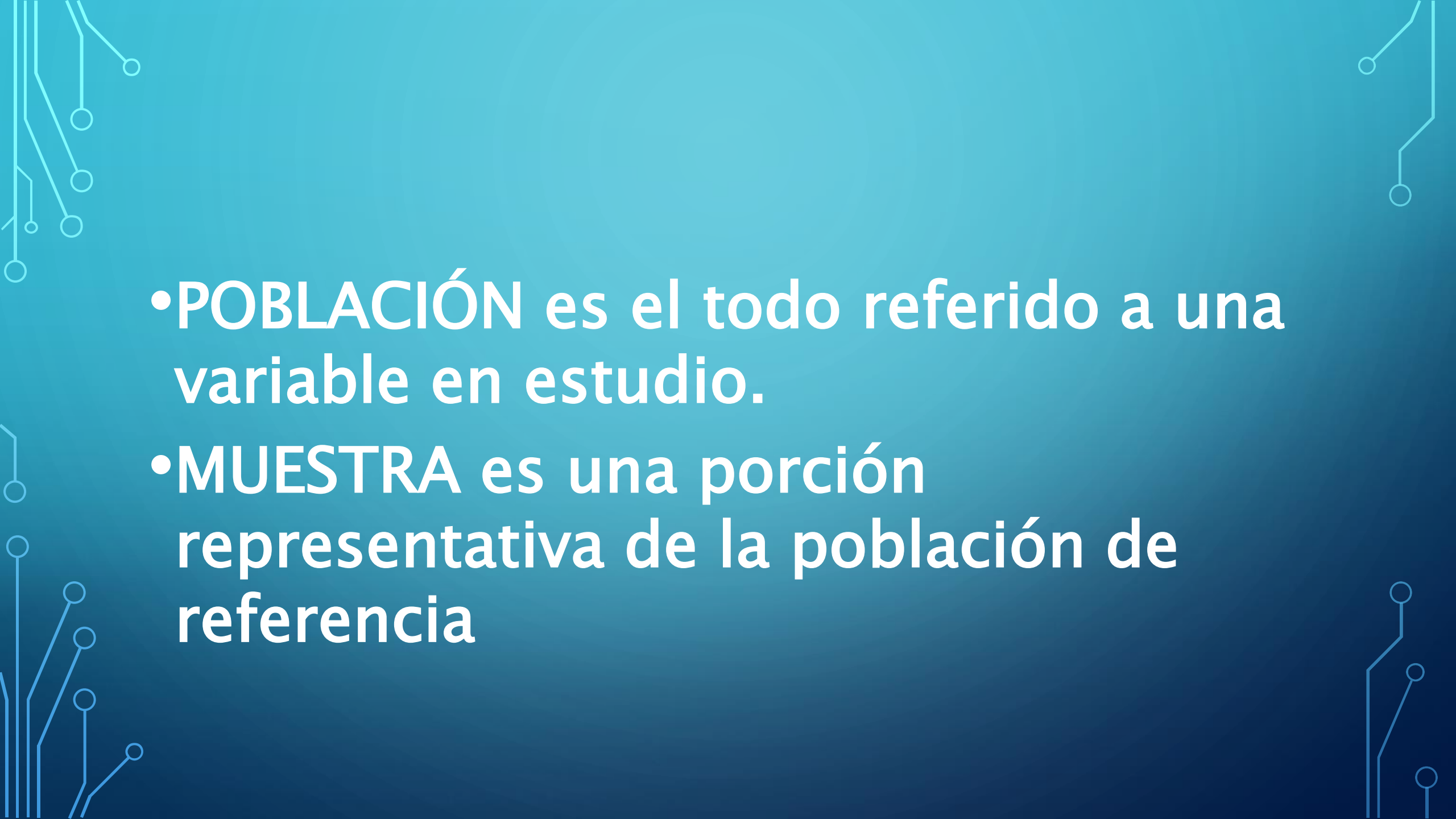
The background is a blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or data paths, with small circles at the end of the lines.

# ESTADÍSTICA DESCRIPTIVA

## ANÁLISIS Y EXPLORACIÓN DE DATOS

# ESTADÍSTICA DESCRIPTIVA

- Es la parte de la Estadística que realiza una descripción numérica, ordenada y simplificada, con la ayuda de representaciones gráficas, de la información obtenida en el relevamiento de datos de una situación en estudio.

- 
- The background is a solid blue gradient. It is decorated with white, stylized circuit board traces. These traces are located in the corners and along the edges, featuring small circles at various points, resembling electronic components or nodes in a network.
- **POBLACIÓN** es el todo referido a una variable en estudio.
  - **MUESTRA** es una porción representativa de la población de referencia

- **Caracteres estadísticos:** es una propiedad que permite clasificar a los elementos de una población.
- *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

# DATOS CUANTITATIVOS

- En el ordenamiento de los datos se debe hacer la distinción entre datos (variables) de tipo continuo y discreto.
- La forma de la distribución de los datos (observaciones de una variable) se denomina *distribución de frecuencias*.

- En una población de  $N$  elementos, descrita según una variable o carácter  $X$ , cuyas modalidades han sido agrupadas se define:
- *Frecuencia absoluta*  $f_i$  : número de observaciones o sea el número de veces que se repite el valor  $x_i$ .

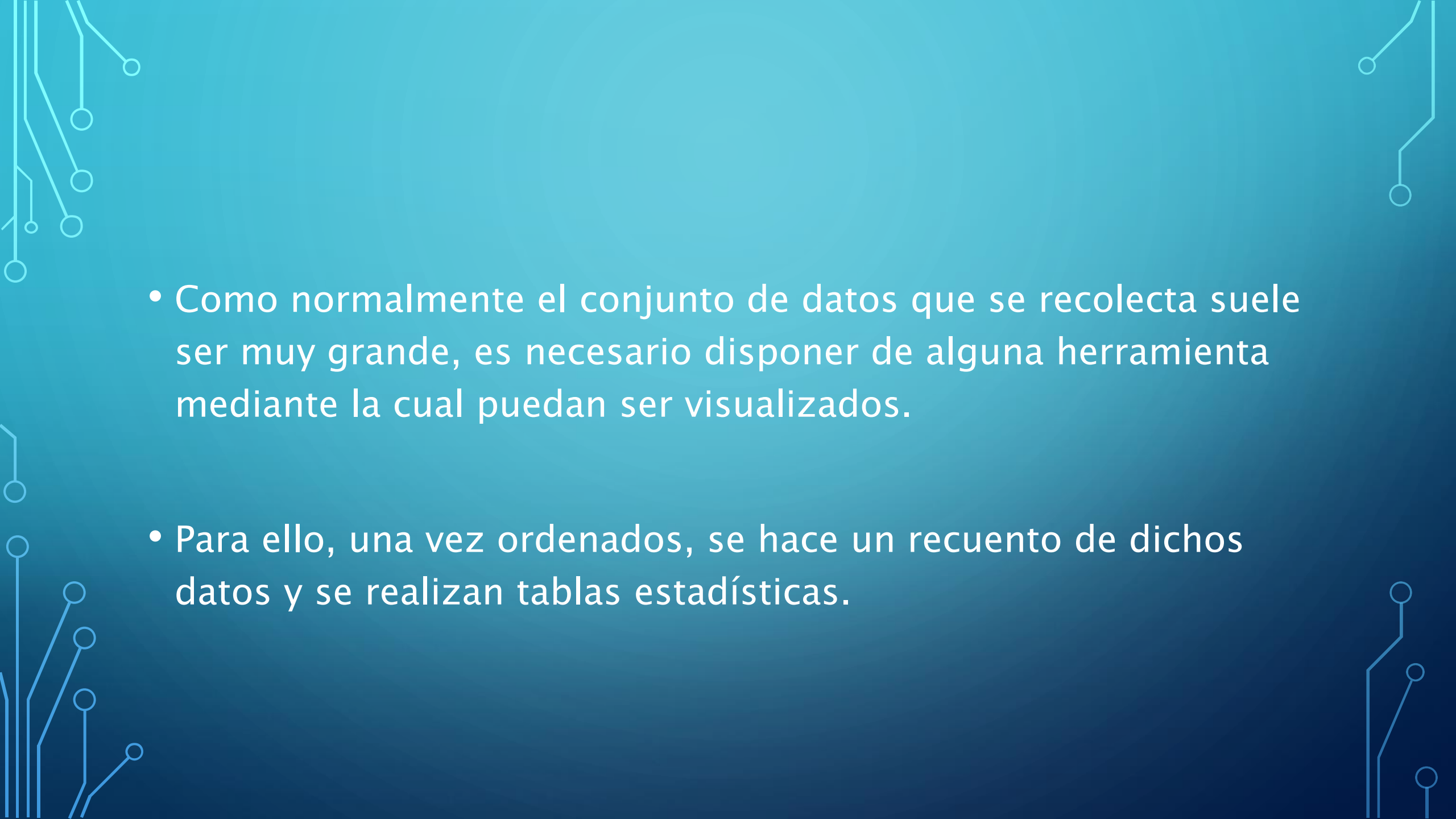


- *Frecuencia absoluta acumulada*  $F_i$  : Es el número de elementos de la muestra cuya modalidad es inferior o equivalente al valor de la variable considerada
- *Frecuencia relativa*  $h_i$  : Es el cociente entre las frecuencias absolutas y el número total de observaciones o datos  $N$

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

- *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:

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

- 
- The background is a solid blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or data paths, with small circles at the end of the lines.
- Como normalmente el conjunto de datos que se recolecta suele ser muy grande, es necesario disponer de alguna herramienta mediante la cual puedan ser visualizados.
  - Para ello, una vez ordenados, se hace un recuento de dichos datos y se realizan tablas estadísticas.

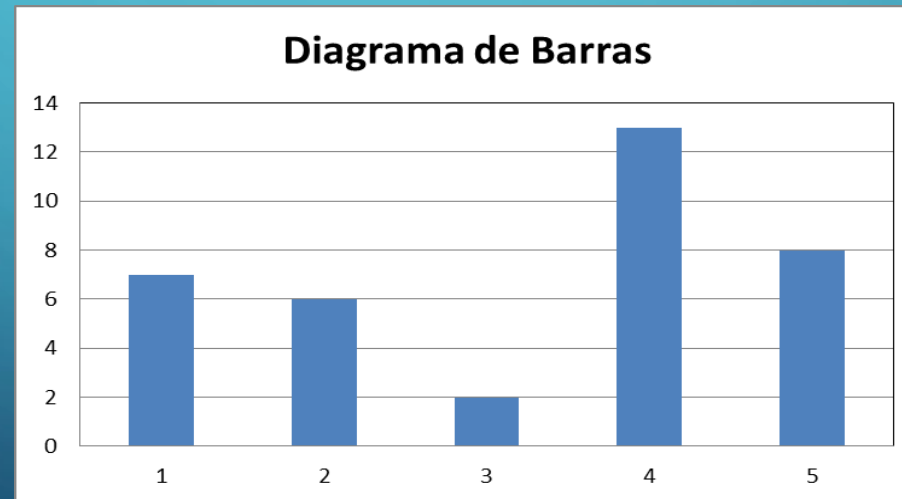
- En estas tablas, deberán figurar los valores de la variable en estudio, y sus frecuencias correspondientes.
- Si bien este ordenamiento hoy es evitable, ya que al trabajar con programas específicos o alguno que posea este tipo de análisis, es útil poder realizarlo para la obtención de algunos gráficos.

- 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 **cuantitativa** se deberá considerar el tipo:
- Si es **discreta**, las modalidades coincidirán con los distintos valores medidos en la muestra.
- Si la variable es **continua** (o a veces discreta, pero toma muchos valores distintos), se tomarán como modalidades **intervalos de clase**. Los intervalos donde se encuentran los datos agrupados, se suelen simbolizar por  $[L_{i-1}, L_i)$ .

# GRÁFICOS

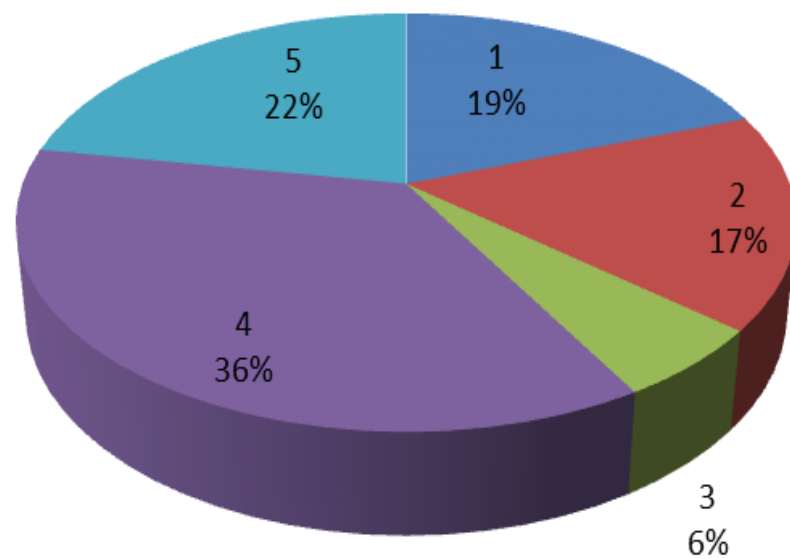
- Gráficos para variables cualitativas
  - Diagrama de barras o bastones



- **Diagramas de sectores o de torta:**
- Se utilizan para hacer comparaciones de las distintas modalidades de un carácter mediante sectores circulares



## Diagrama de Sectores

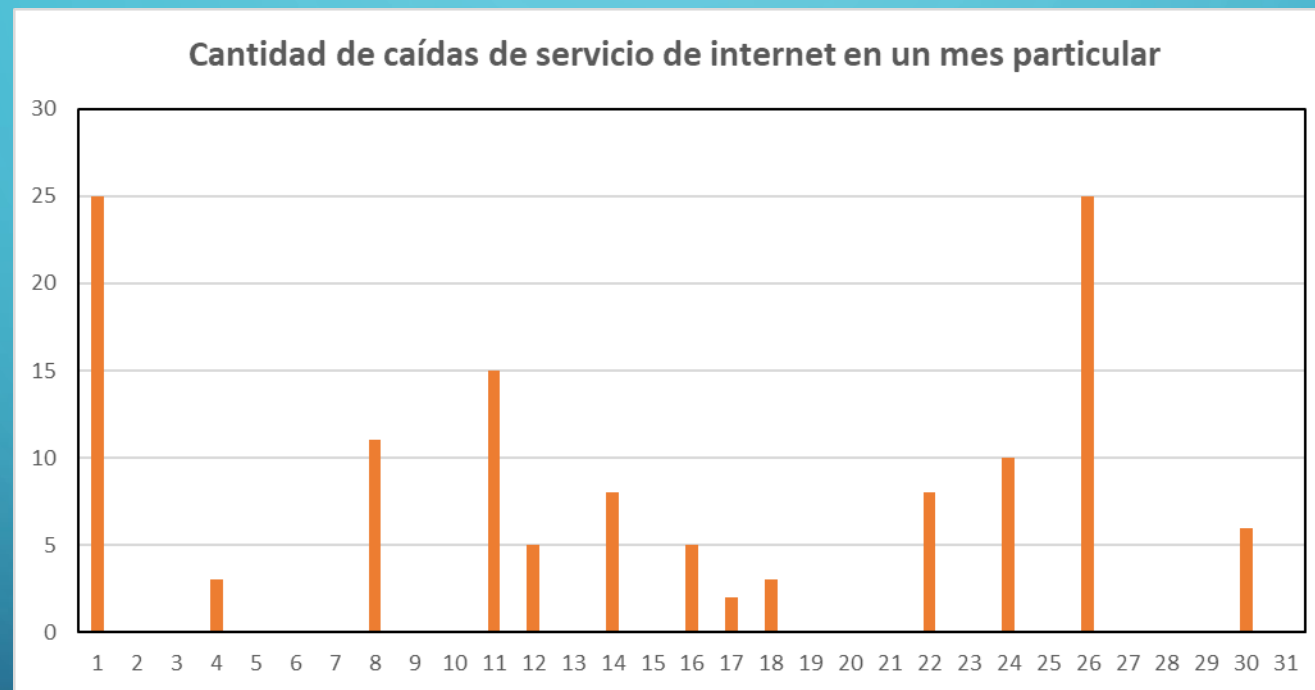


# GRÁFICOS PARA VARIABLES CUANTITATIVAS

- Variables discretas
  - Diagrama de barras, Box Plot
- Variables continuas
  - Histograma, de Tallo y Hojas, Box Plot

# DATOS DISCRETOS

# DIAGRAMA DE BARRAS



# DATOS CONTINUOS



# DISTRIBUCIÓN DE FRECUENCIA

- Los tres pasos necesarios para construir una distribución de frecuencias y definir las clases de la misma con datos cuantitativos son:
  - 1. Determinar el número de clases disyuntas.
  - 2. Determinar el ancho de cada clase
  - 3. Determinar los límites de clase.

- **Número de clases** Las clases se forman especificando los intervalos que se usarán para agrupar los datos.
- Se recomienda emplear entre 5 y 20 clases. Cuando los datos son pocos, cinco o seis clases bastan para resumirlos. Si son muchos, se suele requerir más clases. La idea es tener las clases suficientes para que se muestre la variación en los datos, pero no deben ser demasiadas si algunas de ellas contienen pocos datos.
- **Ancho de clase** El segundo paso al construir una distribución de frecuencia para datos cuantitativos es elegir el ancho de las clases. Como regla general es recomendable que el ancho sea el mismo para todas las clases.

- Como sugerencia se dan dos expresiones para obtener el ancho de clase o el número de clases.
- Es indicativo ya que los softwares pueden hacerlo con valores preestablecidos que pueden ser cambiados según se quiera mejorar la distribución empírica obtenida




$$\frac{\text{Valor mayor en la muestra} - \text{Valor menor en la muestra}}{N^{\circ} \text{ de clases}} = \text{ancho de clase}$$





$$\frac{\text{Valor mayor en la muestra} - \text{Valor menor en la muestra}}{\text{ancho de clases}} = N^{\circ} \text{ de clases}$$

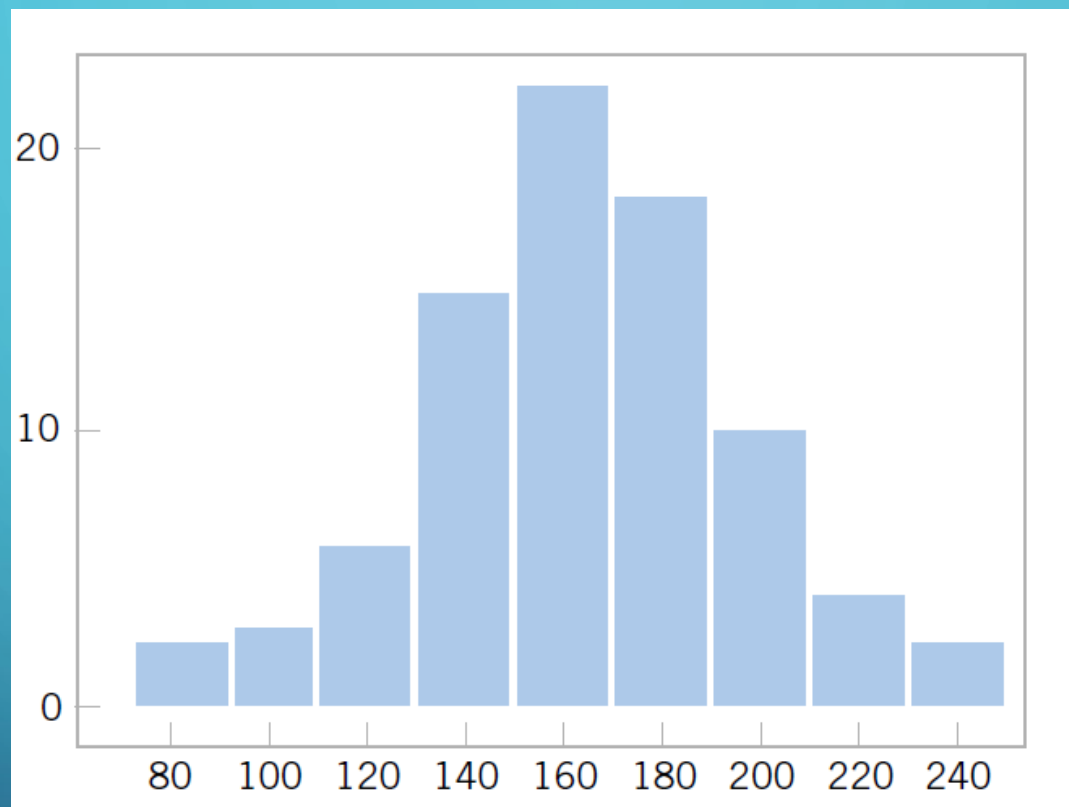
- Al elaborar distribuciones de frecuencia para datos cualitativos, no es necesario especificar límites de clase porque cada dato corresponde de manera natural a una de las clases disyuntas. Pero con datos cuantitativos los límites de clase son necesarios para determinar dónde colocar cada dato.

# HISTOGRAMA

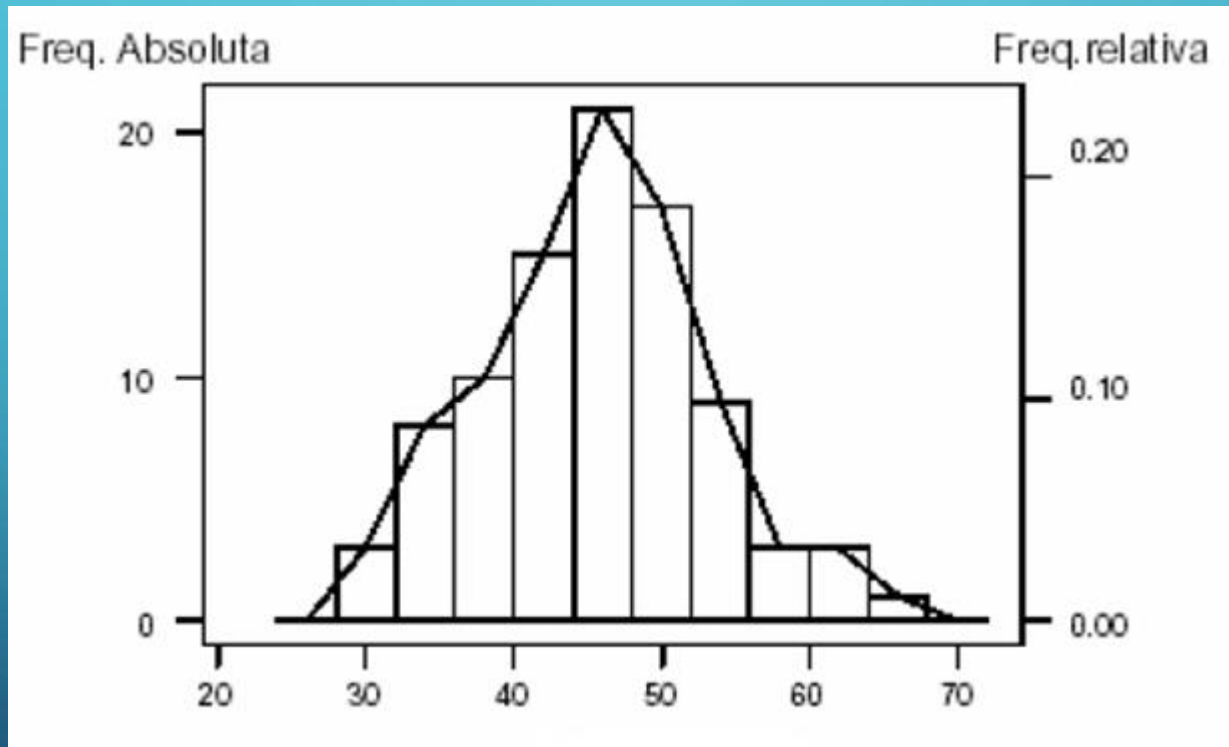


## • Consejos:

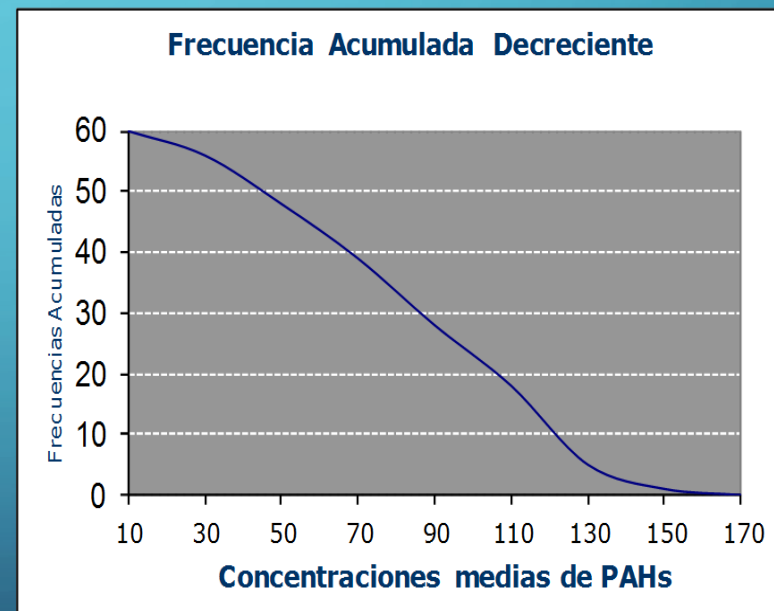
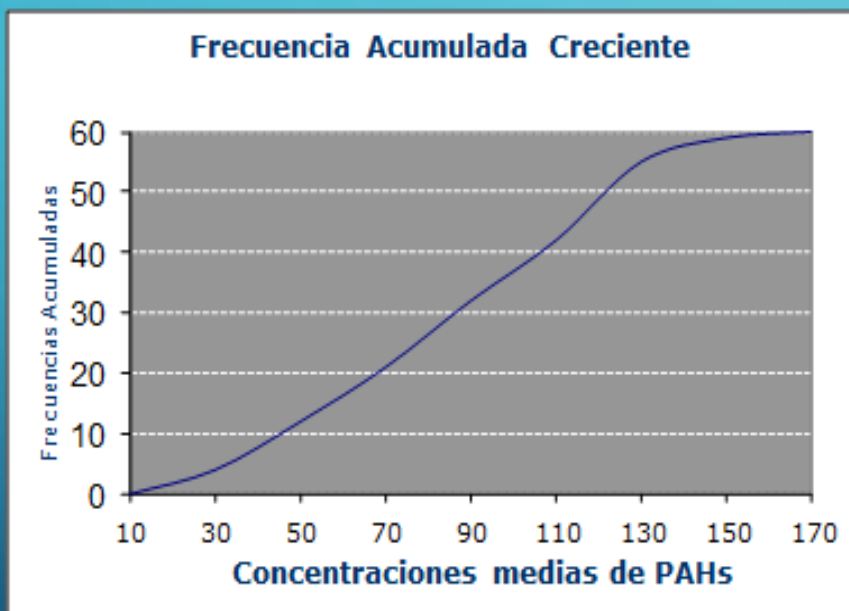
- 1. Usar intervalos de la misma longitud
  - 2. Los intervalos no pueden solaparse
  - 3. Cada observación sólo puede pertenecer a un intervalo
  - 4. Todos los datos deben pertenecer a algún intervalo
- 
- 
- 

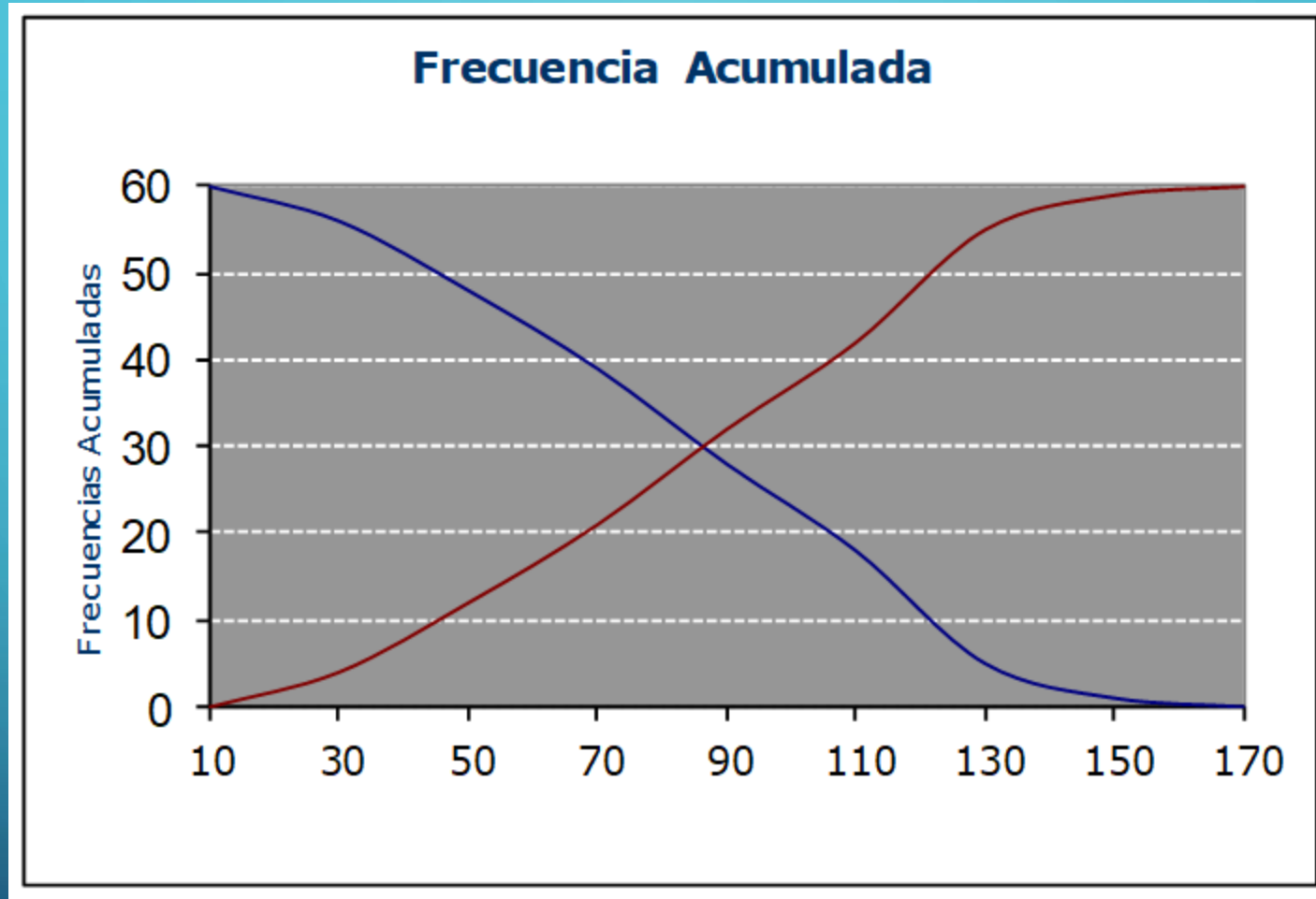


# POLÍGONO DE FRECUENCIAS



# GRÁFICOS DE FRECUENCIAS ACUMULADAS







# ANÁLISIS EXPLORATORIO

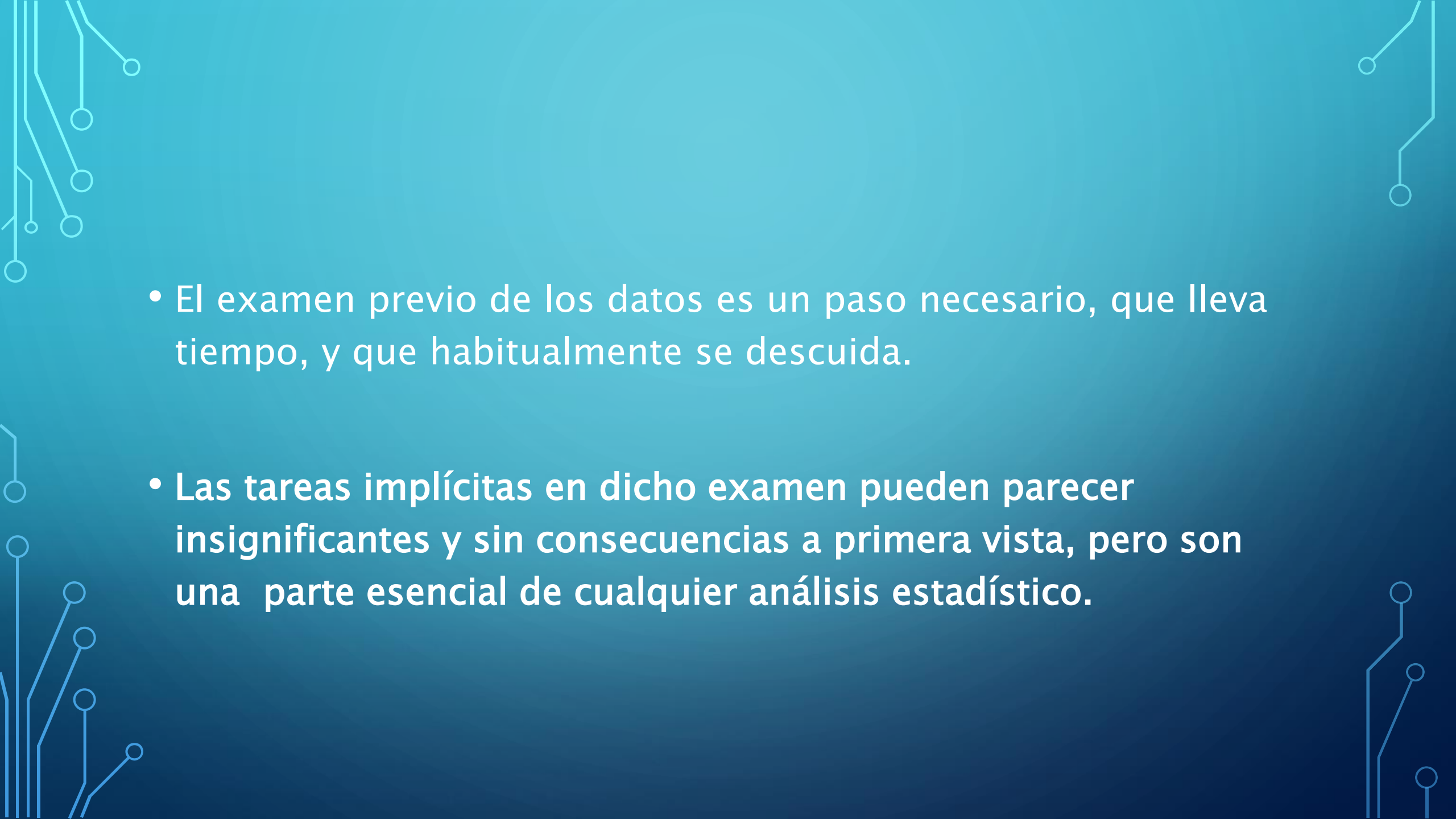
- Análisis reciente, son métodos innovadores para el análisis de datos.
- Hace énfasis en la exploración de los datos por métodos gráficos previos al clásico análisis estadístico.

- 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.
- 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 de los datos.
  - Proporcionar hipótesis o modelos acerca del comportamiento de los datos
  - Se ha robustecido con la reciente aparición de diversos programas específicos con licencia y software libre.



## • Gráficos más importantes :

- – El diagrama de Tallo y Hoja.
  - -El diagrama de Caja y Bigotes.
- 
- 

- 
- The background is a solid blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles connecting them.
- 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.

# Cómo construirlo:

- 1.-Ordenar el lote de datos en magnitud creciente.
- 2. Dividir en dos partes cada 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. Describe 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 tratada.





0|99  
1|001111222223333334  
44444  
1|556778  
2|011222334  
2|677888899  
3|00111122234  
3|5568899  
4|011122224444  
4|55566677788888  
5|0012



1	0 5
7	0 666777
26	0 88899999999999999999
(20)	1 00000000001111111111
43	1 2233333
36	1 444444444444555555
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



# CARACTERÍSTICAS

- Son números que sirven para obtener valores resúmenes para el estudio de la **muestra**.
- Se dividen en grupo según lo que permiten estudiar



• **MEDIDAS DE TENDENCIA CENTRAL**

• **MEDIDAS DE DISPERSIÓN**

• **MEDIDAS DE FORMA:**

- ***ASIMETRIA***

- ***CURTOSIS***

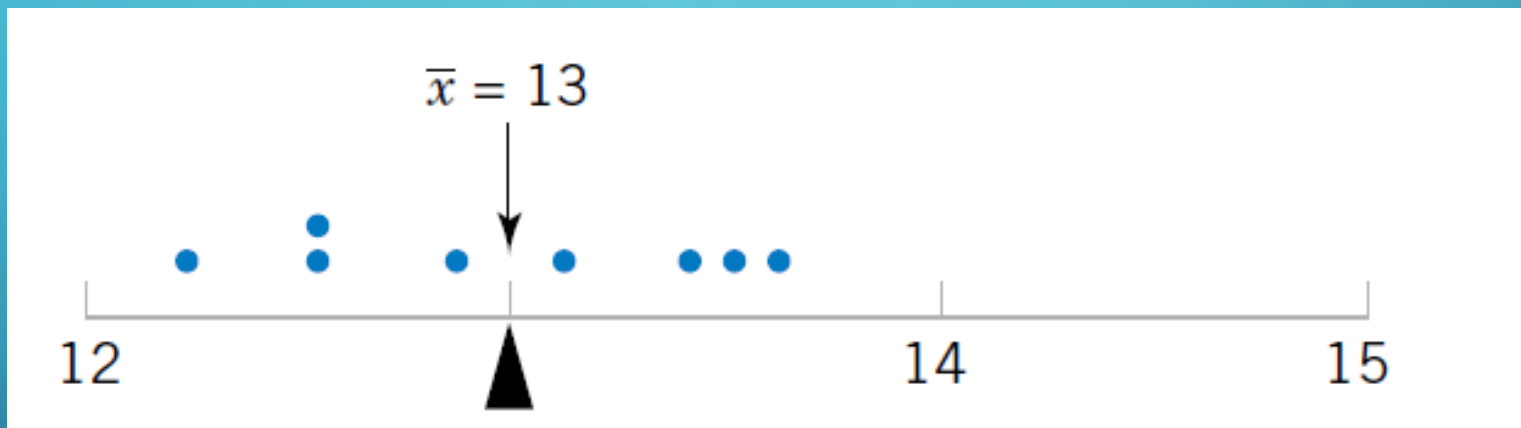


# MEDIDAS DE TENDENCIA CENTRAL

- Promedios

- Media aritmética o media de muestra:

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



# PROPIEDADES

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

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

$$\sum (x_i - \bar{x})^2 = \textit{mínimo}$$

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



# OTROS PROMEDIOS:

- Media Geométrica

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

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

- Media Armónica:

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

# MEDIDAS DE UBICACIÓN

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

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

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

# MEDIANA

- Variables discretas:
- Si no hay frecuencias
  - – Número de datos impar: la Mna. es el valor central.
  - – Número de datos par: la Mna. es el promedio de los valores centrales.

- Si hay frecuencias:
- - Se obtienen las frecuencias acumuladas ( $N_i$ ) y se calcula  $N/2$ :

- *Se distinguen dos casos:*

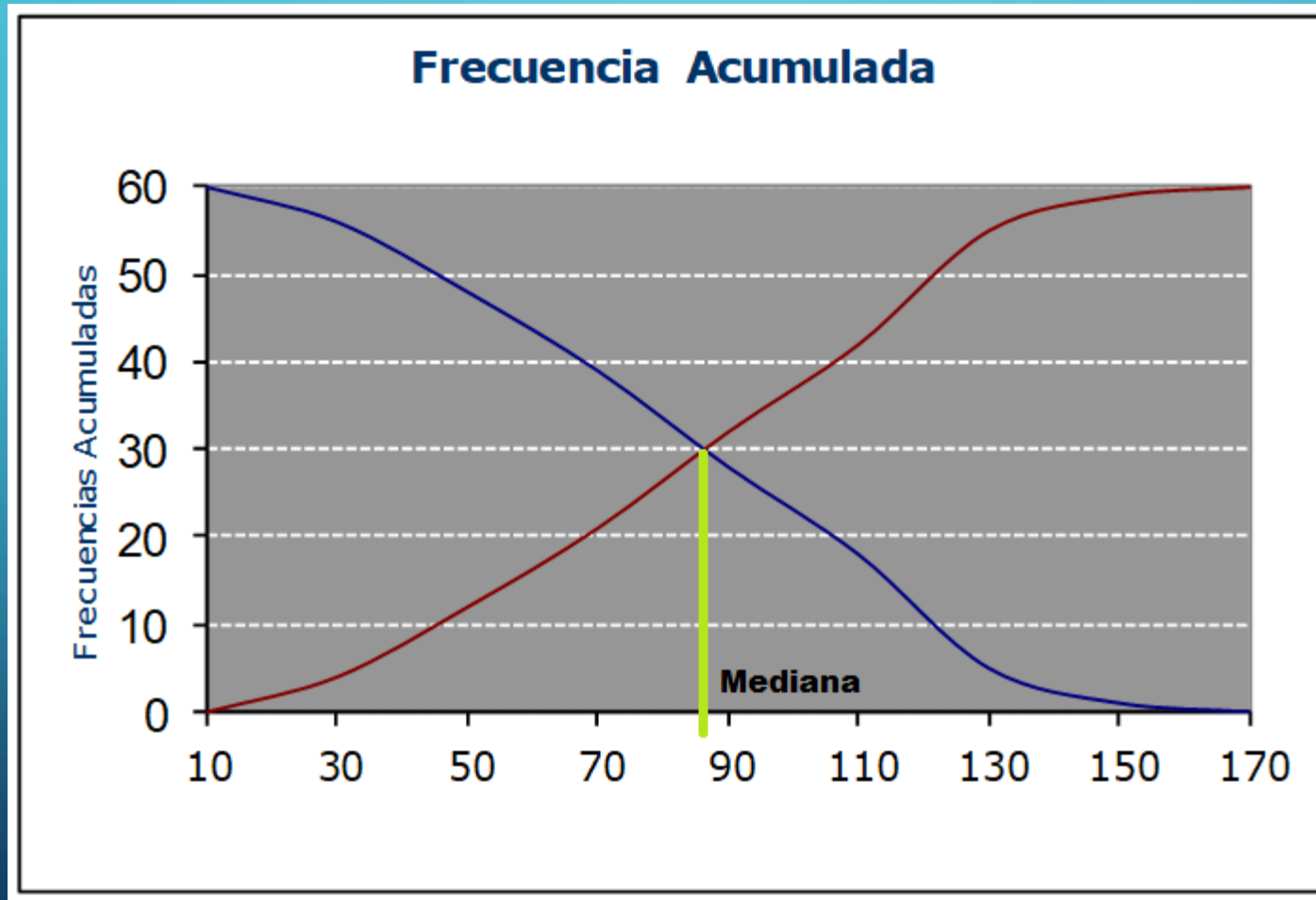
- - Si existe un valor tal que  $N_{i-1}$  **es un número decimal** entonces se toma la ***Mna. = al valor de  $x_i$  correspondiente***
- - Si existe un valor tal que  $N_i = N/2$  entonces la mediana será el promedio entre el valor correspondiente de  $x_i$  y el siguiente:

$$Mna. = \frac{x_i + x_{i+1}}{2}$$

- Variables continuas: se obtiene interpolando, hoy se dispone de su cálculo en los softwares.

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

- También se la puede obtener gráficamente:



# PROPIEDADES

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

# CUANTILES

- Como la mediana divide a la distribución de datos en dos partes, los **cuartiles** la dividen en cuatro, los **deciles** en diez y los **percentiles** en cien.
- Se calculan de la misma forma que la mediana, sólo que cambia como se determina el orden del cuantil.



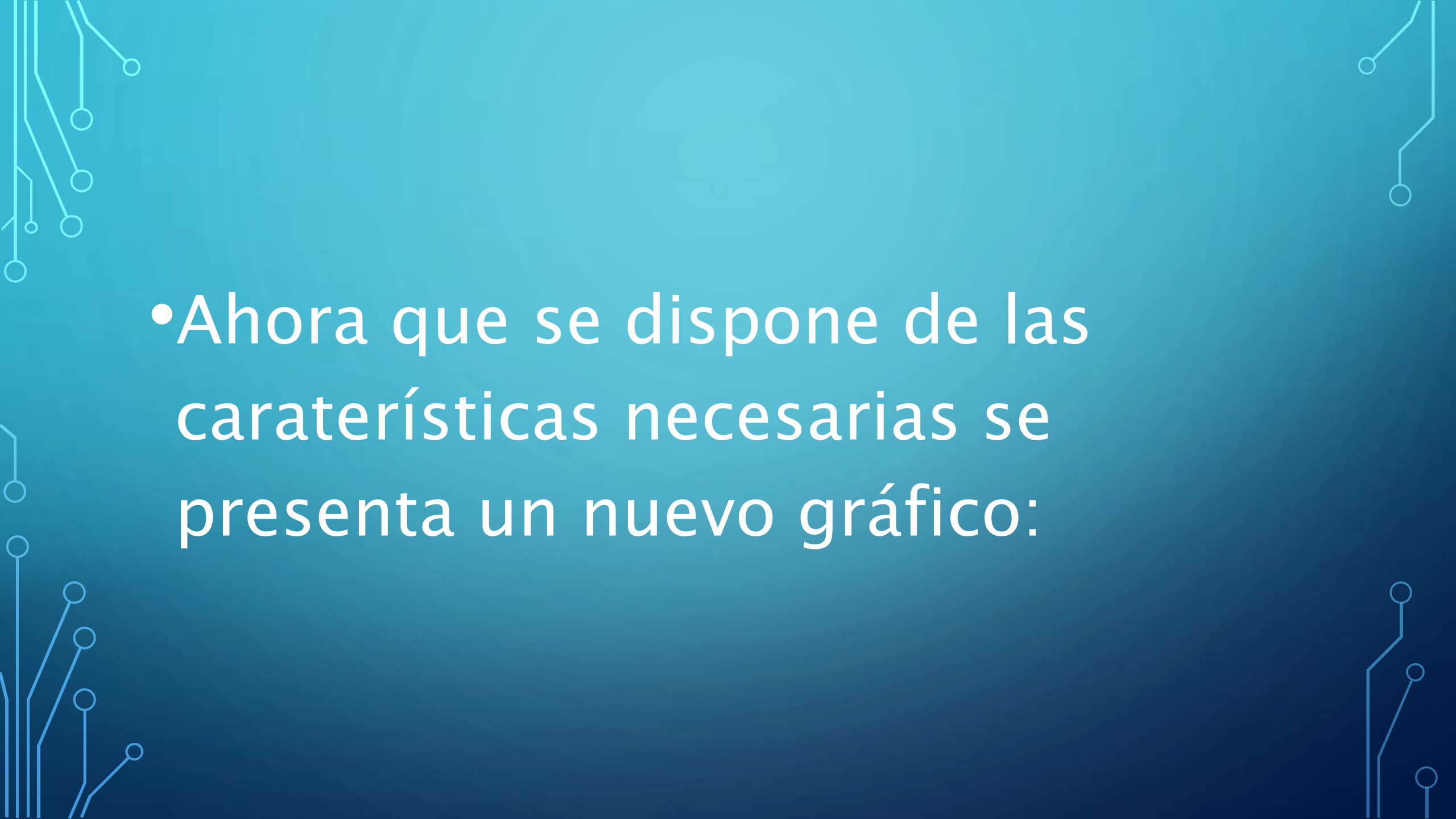
## CÁLCULO DE UN PERCENTIL

- **Paso 1.** Ordenar los datos de menor a mayor (colocar los datos en orden ascendente).
- **Paso 2.** Calcular el orden:

$$i = \left( \frac{p}{100} \right) n$$


donde ***p*** es el percentil que se quiere obtener y ***n*** es el número de observaciones.

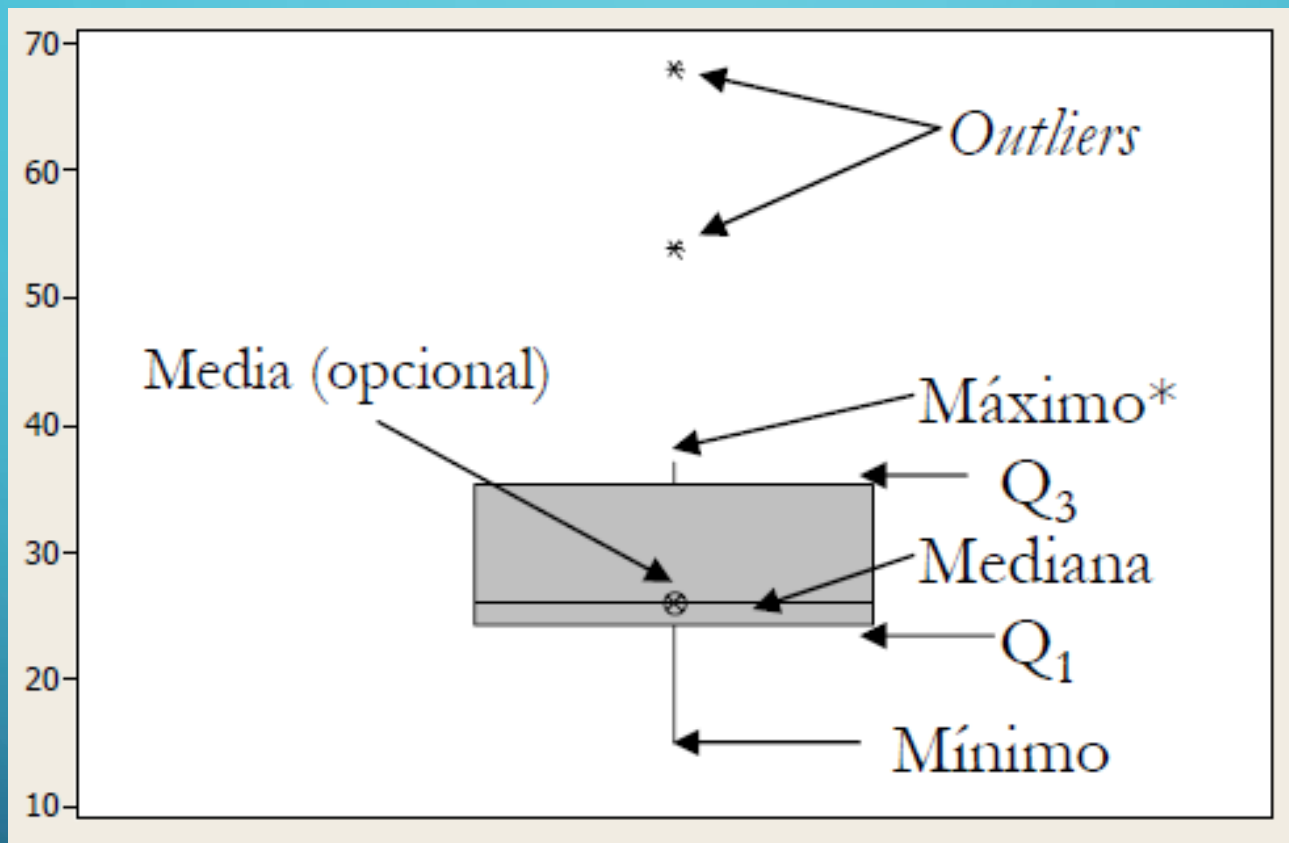
- **Paso 3.**
- (a) Si *i* no es un número entero, se debe redondear. El primer entero mayor que *i* denota la **posición** del percentil ***p***.
- (b) Si *i* es un número entero, el percentil ***p*** es el **promedio** de los valores en las posiciones *i* e *i* + 1.

- 
- The background is a blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles.
- Ahora que se dispone de las características necesarias se presenta un nuevo gráfico:

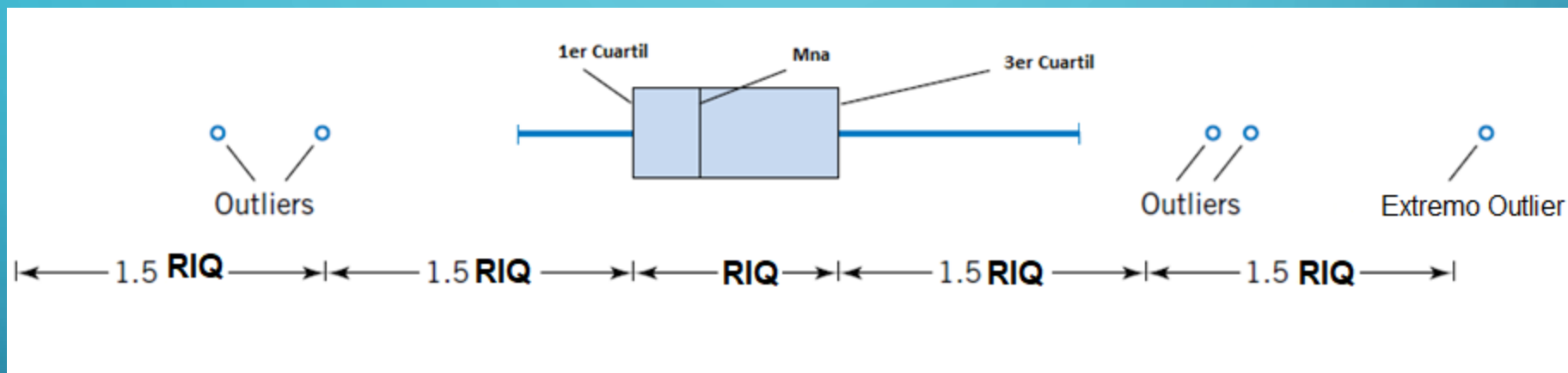
# GRÁFICO DE CAJA Y BIGOTE

- 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.

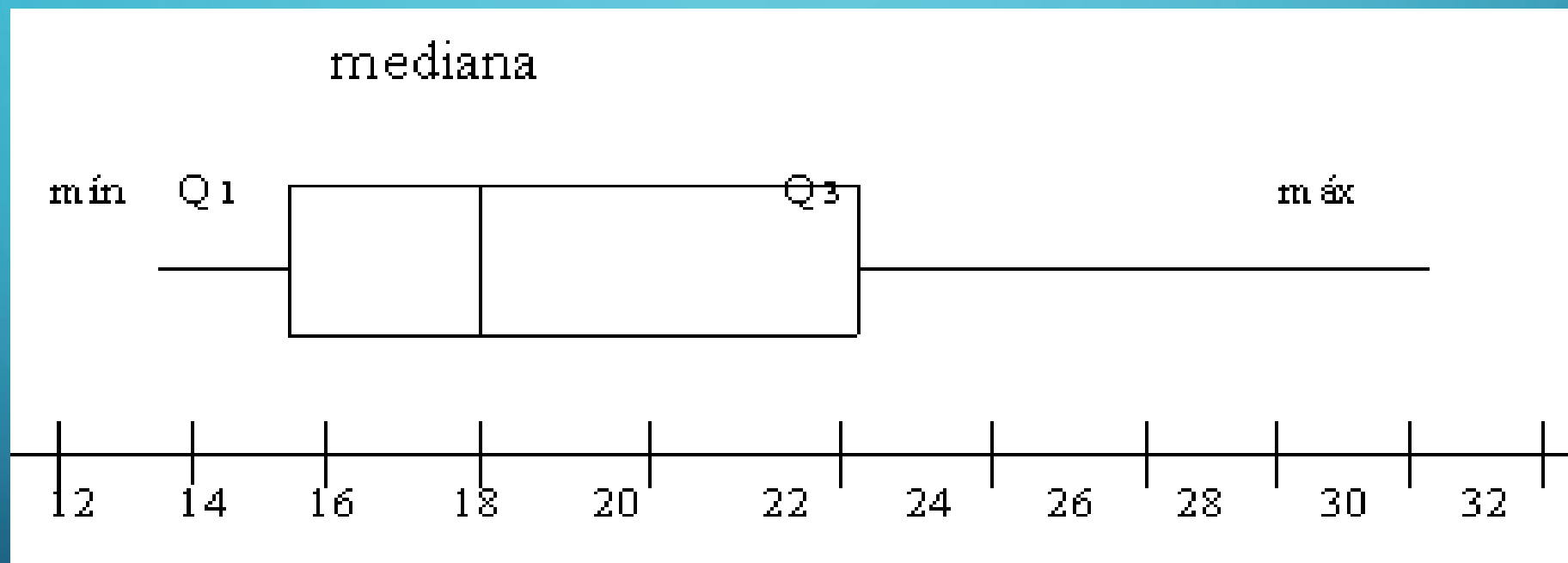
- 
- The slide features a dark blue background with decorative white circuit-like lines in the corners. These lines consist of vertical and horizontal segments connected by small circles, resembling a stylized electronic circuit or data network.
- Es posible introducir algunas variaciones en la construcción 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.

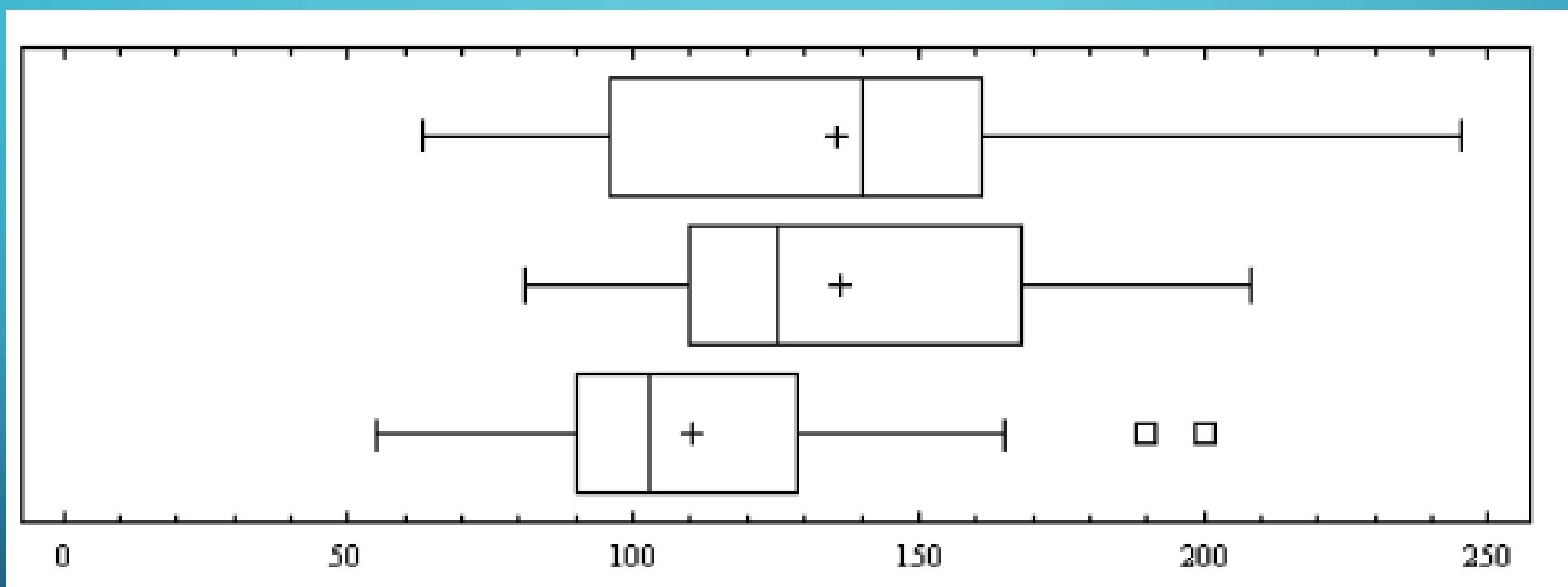


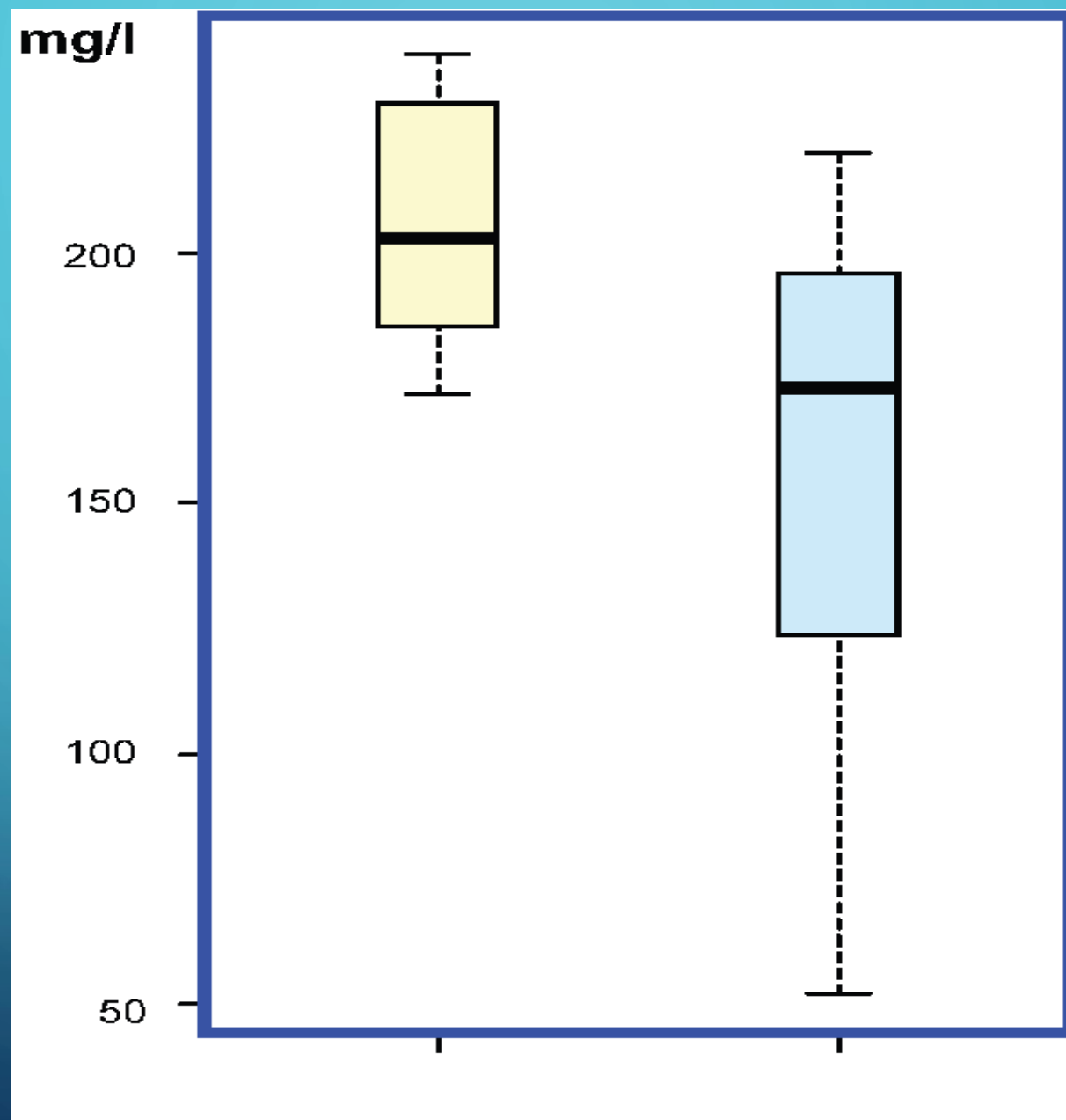


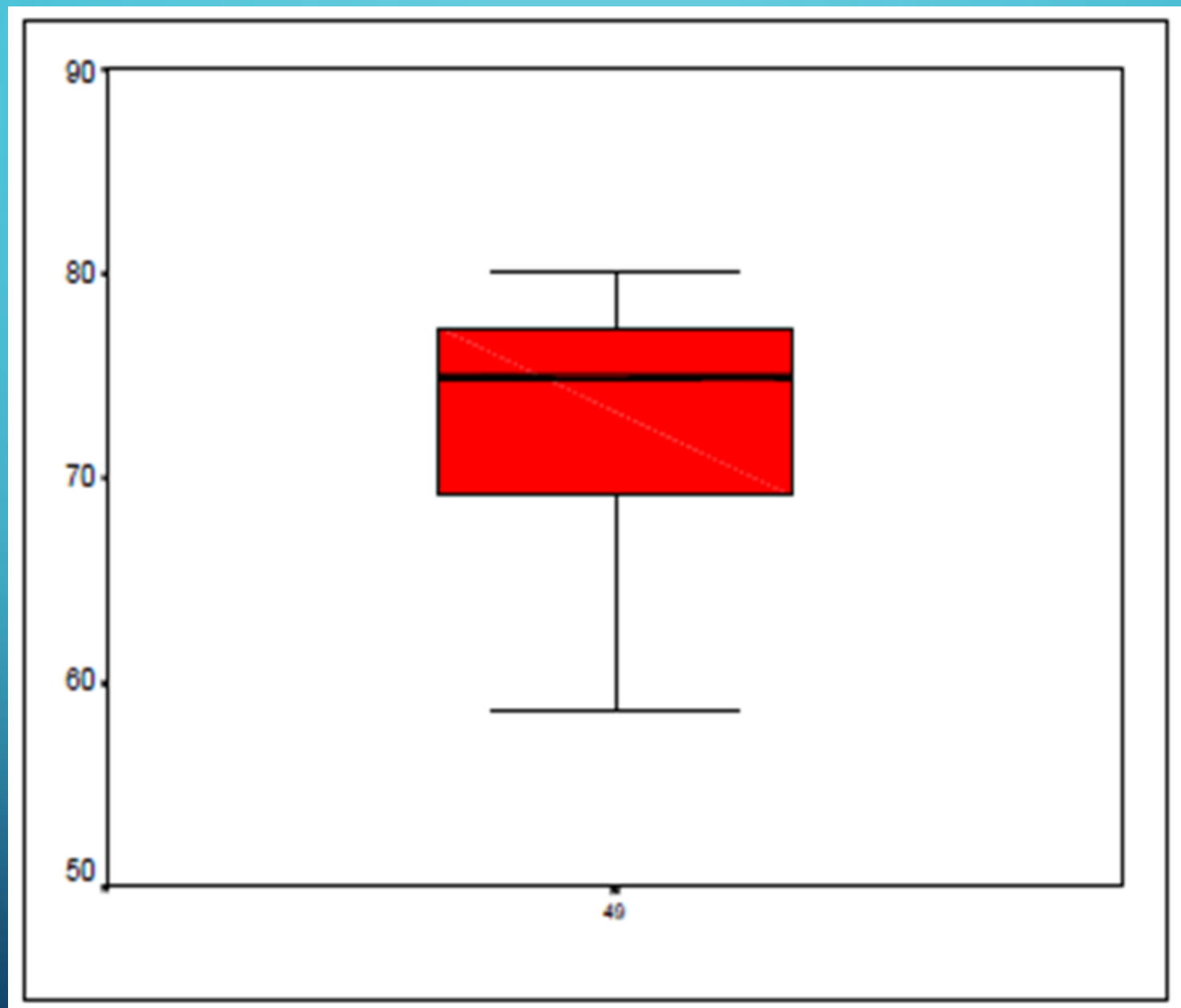


- Puede construirse sin tener en cuenta escala.
  - De forma horizontal o vertical
  - Sirve para comparar la variabilidad y asimetría de varias muestras
- 
- El ancho de la caja esta definido por el rango inter-cuartílico (Q1 y Q3). A mayor amplitud de la caja, mayor variabilidad en los datos.





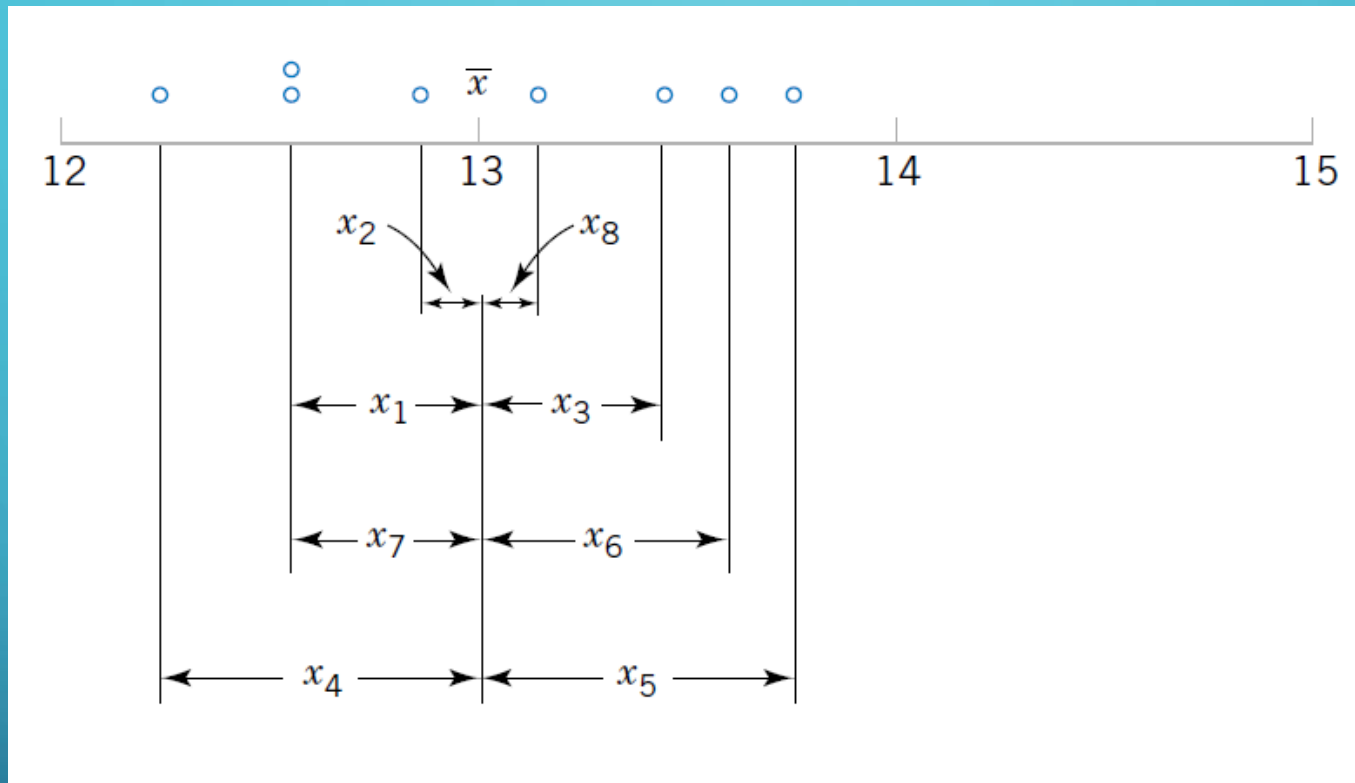




# MEDIDAS DE DISPERSION

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

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

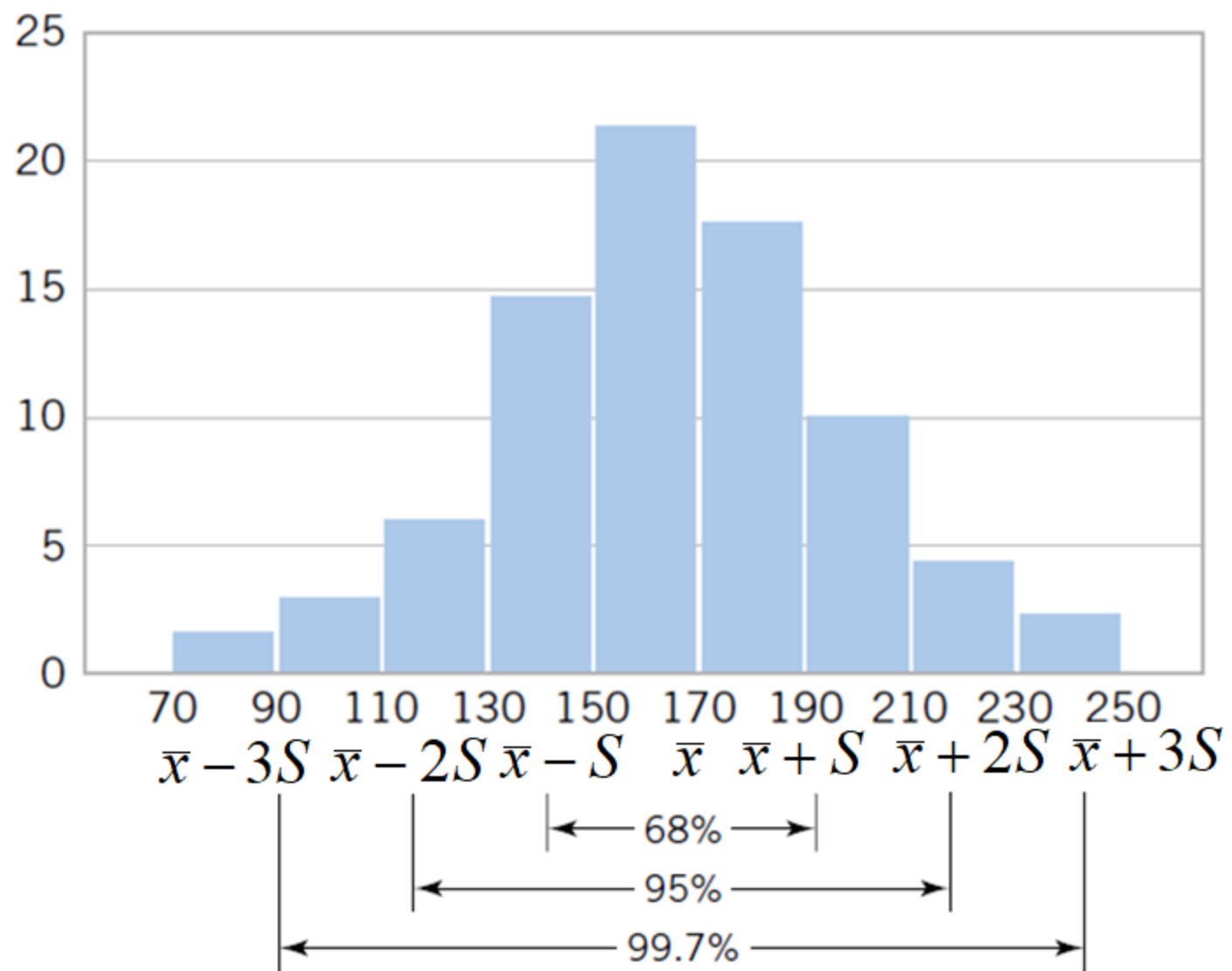




$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$$Cv = \frac{S}{\bar{x}} (\%)$$

- **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 simétrica.**
- **Esta regla considera que:**
  - **si se miden en el eje x 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.**



# MEDIDAS DE FORMA

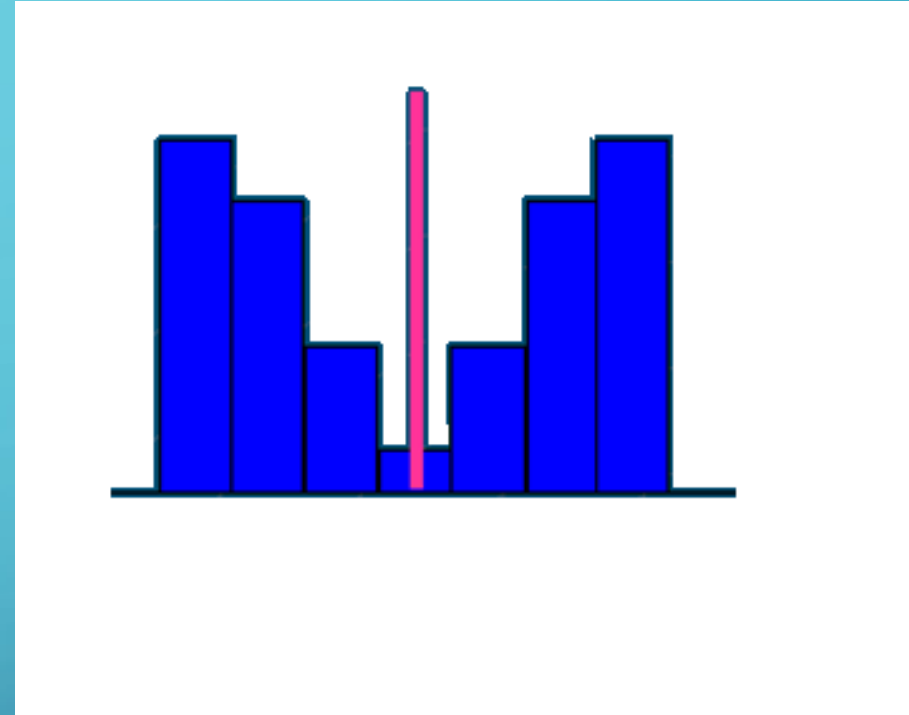
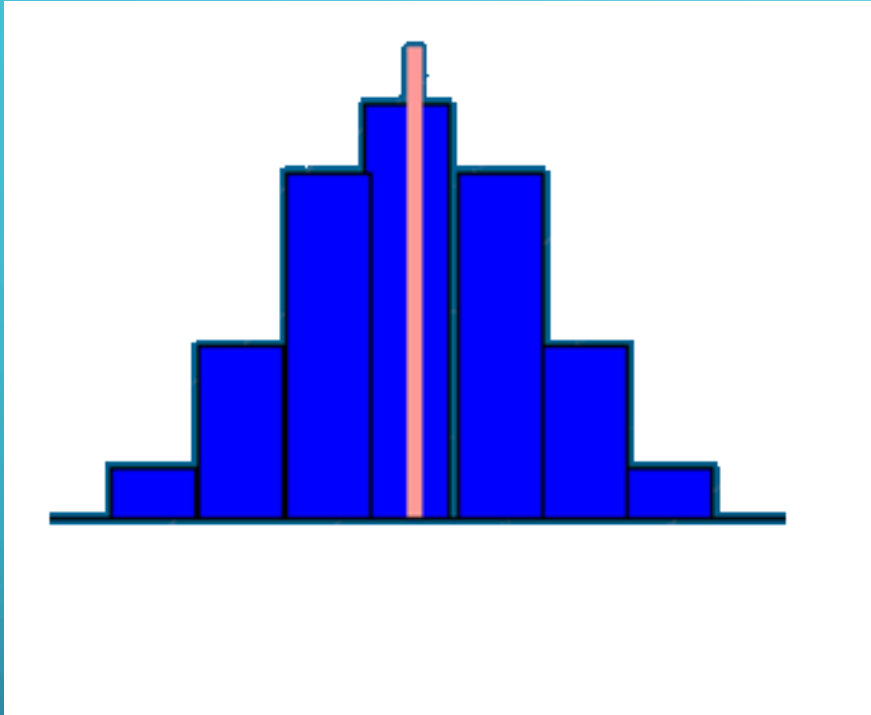
- ASIMETRÍA

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

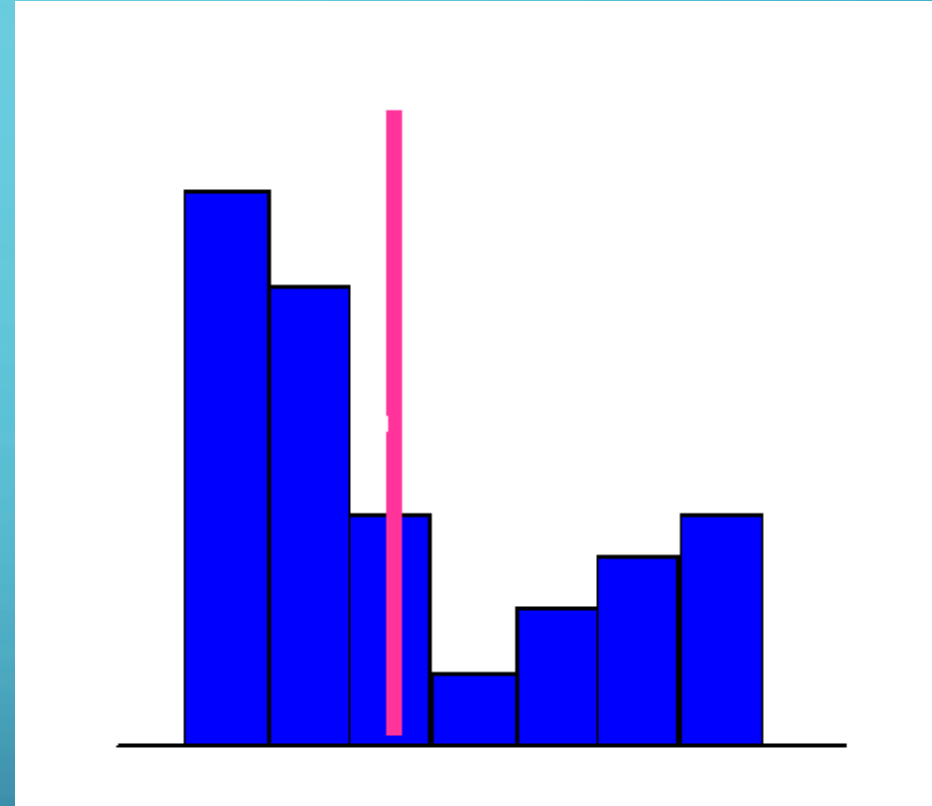
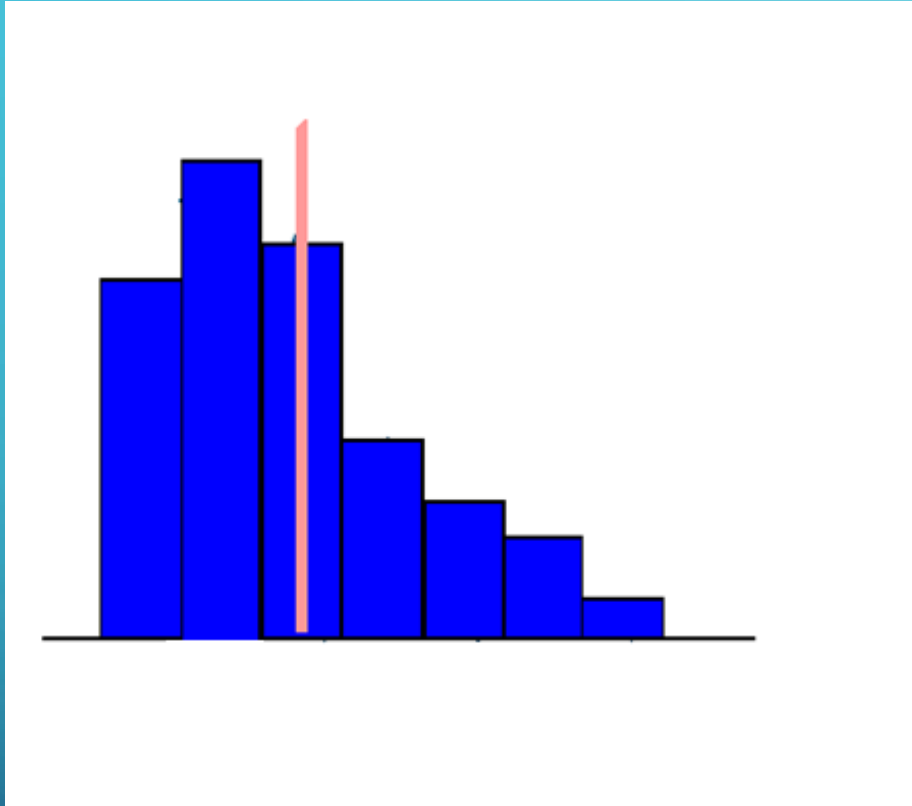
$$As = \frac{3(\bar{x} - \textit{Mediana})}{S}$$

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

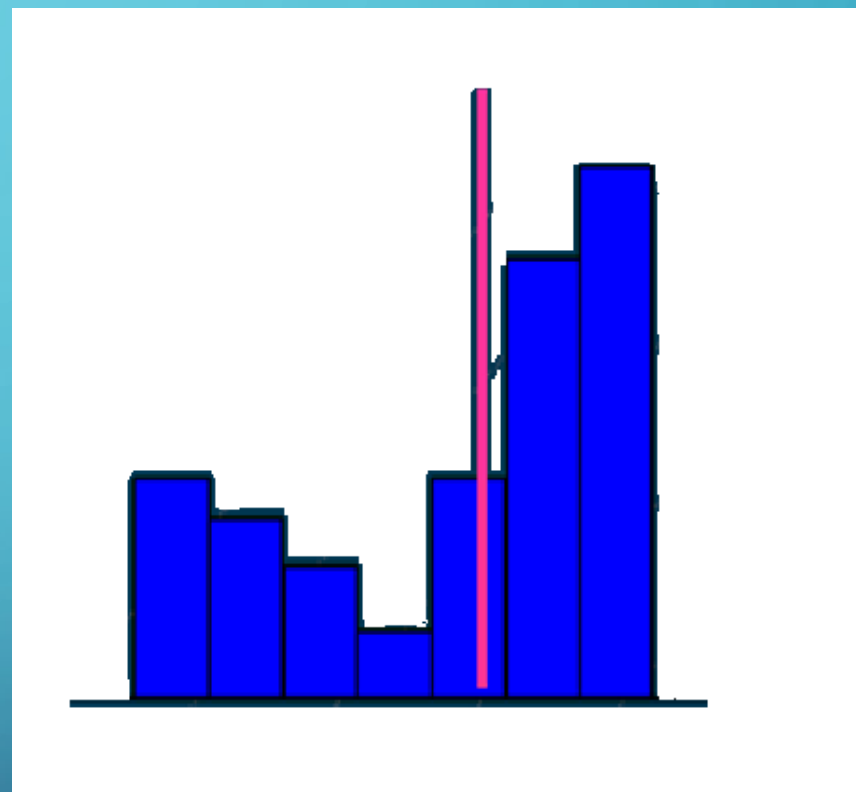
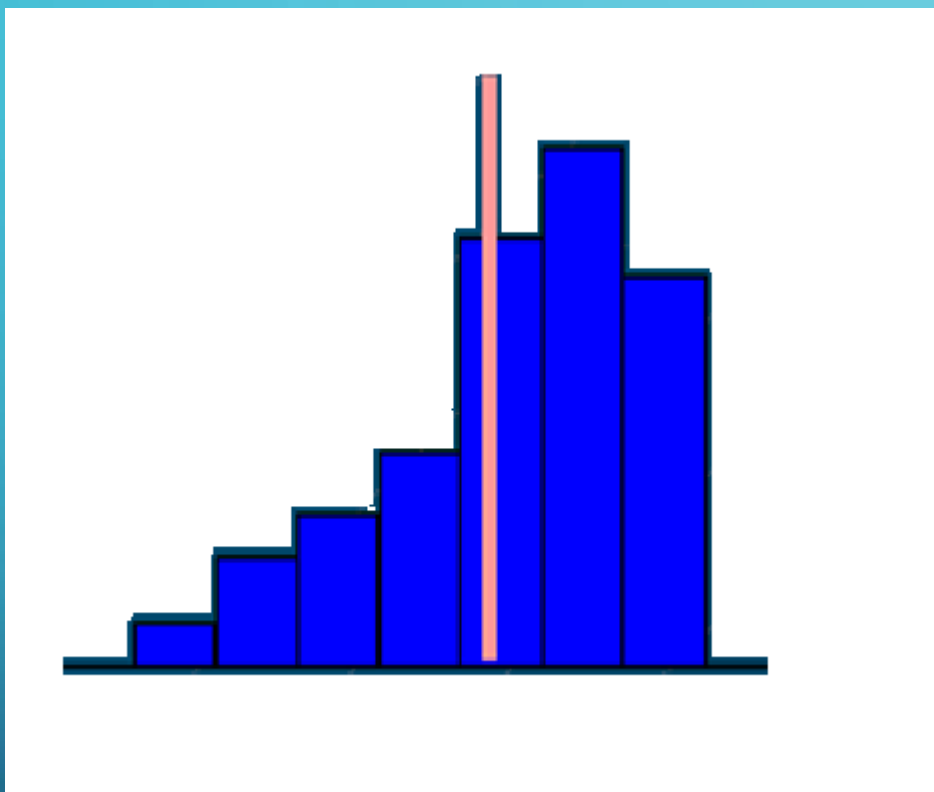
$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n}$$



Distribución simétrica



Distribución asimétrica  
positiva o a la derecha



**Distribución asimétrica  
negativa o a la izquierda**

- CURTOSIS

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

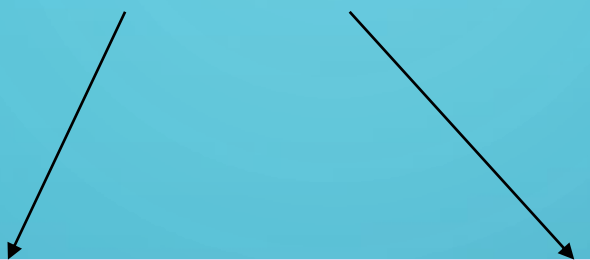
$$K = \frac{m_4}{S^4}$$



## Medidas Descriptivas Numéricas y Representaciones Gráficas 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	Coefficiente de Variación

- Notaciones en **MUESTRA** y **POBLACIÓN**



	Estadístico	Parámetro
Media	$\bar{x}$	$\mu$
Varianza	$s^2$	$\sigma^2$
Desviación estándar	$s$	$\sigma$
Covarianza	$s_{xy}$	$\sigma_{xy}$
Correlación	$r_{xy}$	$\rho_{xy}$

# PROPUESTA

- Los invitamos a buscar información que sea de interés particular y aplicar todos los conceptos aprendidos hasta acá en un trabajo personal o en equipo de no más de dos personas.