

## Presentación de los datos

### Presentación Escrita

Este método consiste en presentar un informe que reseña los rasgos de mayor importancia de los datos. Debido a que es necesario leer el informe íntegramente para conocer los aspectos de interés de los datos, este método no es muy efectivo y por consiguiente es poco empleado. Sin embargo, posee la virtud de poder resaltar las cifras y las comparaciones que se consideren esenciales.

### Ejemplo:

En el boletín mensual sobre el *índice nacional de precios al consumidor* (INPC) presentado por el Banco Central de Venezuela, en fecha 03/05/2012, se puede leer:

### **INDICE NACIONAL DE PRECIOS AL CONSUMIDOR PARA EL MES DE ABRIL DE 2012**

“El Índice Nacional de Precios al Consumidor (INPC), elaborado por el Banco Central de Venezuela (BCV) y el Instituto Nacional de Estadística (INE), arrojó en el mes de abril de 2012 una variación intermensual de 0,8%, la más baja desde que se comenzó a calcular el indicador con cobertura nacional (desde enero de 2008). Con este registro se logra la quinta desaceleración consecutiva de la tasa mensual, secuencia que ha permitido mejorar desde 2,2% en noviembre de 2011 a 0,8% en el recién finalizado mes de abril.

Al cierre del primer cuatrimestre del año la variación acumulada se ubica en 4,4%, significativamente por debajo del 7,5% que se tenía en similar período del año anterior. Por su parte, la variación anualizada se ubicó en 23,8%, con lo que desacelera por cuarto mes consecutivo, desde el 27,6% obtenido en diciembre de 2011.

Al evaluar los resultados por categorías de rubros se tiene que 2 agrupaciones mostraron una variación intermensual negativa: Equipamiento del hogar, con -3,0%, y Bienes y servicios diversos, con -2,4%. Estos resultados reflejan el efecto de la providencia administrativa emitida por la Superintendencia Nacional de Costos y Precios Justos (Sundecop), publicada en Gaceta Oficial N° 39.871, de fecha 27 de febrero de 2012, y con vigencia a partir del 1° de abril. En dicha disposición legal se fija el precio de 19 productos, principalmente en artículos para el mantenimiento de la vivienda, de cuidado personal y en algunos alimentos. A estas 2 agrupaciones con variación negativa se añaden otras 3 que, aun siendo positivas, se encuentran muy por debajo del promedio global (0,8%): Vestido y calzado (0,2%), Servicio de la vivienda (0,3%) y Comunicaciones (0,4%). Los restantes 8 grupos superaron el promedio, con un máximo de 3,5% en el conjunto de las Bebidas alcohólicas y tabaco.

Desde el punto de vista de la dinámica intermensual ocurre que 6 grupos desaceleran, a la vez que 7, por el contrario, presentan una mayor variación intermensual en abril. Los que desaceleran

son: Equipamiento del hogar (de 1,2% a -3,0%), Bienes y servicios diversos (de 0,8% a -2,4%), Vestido y calzado (de 0,6% a 0,2%), Servicios de la vivienda (de 0,7% a 0,3%), Salud (de 1,9% a 1,8%) y Restaurantes y hoteles (de 1,7% a 1,6%). Las aceleraciones, por su parte, se observan en: Alquiler de vivienda (de 0,9% a 1,0%), Transporte (de 0,8% a 1,0%), Comunicaciones (de 0,2% a 0,4%), Servicios de educación (de 1,3% a 1,8%), Alimentos y bebidas no alcohólicas (de 0,5% a 1,2%), Esparcimiento y cultura (de 1,4% a 2,4%) y Bebidas alcohólicas y tabaco (de 1,4% a 3,5%).

El resultado favorable del INPC en el cuarto mes del año, apreciado en el contexto geográfico, obedece a que 7 de los 11 dominios de estudio presentaron un menor crecimiento en abril, en comparación con el mes de marzo: Valencia (de 1,1% a 0,6%), Barquisimeto (de 1,2% a 0,7%), Maturín (de 0,9% a 0,5%), San Cristóbal (de 1,0% a 0,6%), Mérida (de 1,2% a 0,9%), Maracay (de 0,9% a 0,7%) y Caracas (de 1,0% a 0,9%). En 3 dominios la tasa resultó igual a la del mes precedente: Barcelona-Puerto La Cruz (0,7%), Maracaibo (0,9%) y Resto Nacional (0,8%). La única ciudad con evolución adversa fue Ciudad Guayana, donde el crecimiento de abril fue superior al de marzo (de 0,6% pasó a 0,9%).

La subdivisión de la canasta del INPC en bienes y servicios revela comportamientos contrapuestos: mientras los bienes desaceleraron, de 0,6% a 0,4%, los servicios aceleraron, de 1,4% a 1,5%; se mantiene así la tendencia a un menor crecimiento de los bienes iniciada en enero del presente año.

La variación intermensual del núcleo inflacionario aumentó de 1,3% en marzo a 1,6% en abril, lo que responde a sendas aceleraciones ocurridas en Alimentos elaborados (de 1,4% a 2,3%) y Bienes industriales distintos de alimentos y textiles (de 0,9% a 1,7%). En Textiles y prendas de vestir y Servicios no administrados, en cambio, se registraron desaceleraciones (de 0,6% a 0,3% y de 1,7% a 1,6%, respectivamente).

El indicador de escasez repuntó, al pasar de 10,8% a 14,8%, al igual que el de diversidad, que en este caso mejoró de 159,7 a 169,1.”

(Disponible en:

<http://www.bcv.org.ve/c4/notasprensa.asp?Codigo=9890&Operacion=2&Sec=False> [consulta: 2012, Mayo 7] )

## Cuadros Estadísticos

Los cuadros estadísticos son tablas en las cuales se exhibe de manera ordenada a los datos. Un cuadro estadístico debe ser capaz de explicarse por sí solo. Para cumplir esto debe poseer principalmente título, encabezados, cuerpo y fuente.

El *título* debe ser breve y suficientemente explicativo de la situación estudiada, la época y el sitio. *Los encabezados* son los nombres de las filas y columnas de la tabla. *El cuerpo* son los datos ya condensados y organizados. *La fuente* indica el origen de la información, por esta razón nunca debe faltar en todo cuadro estadístico.

Adicionalmente a los elementos anteriores, estos cuadros pueden llevar *notas preliminares* en el título del cuadro, *notas explicativas* debajo del cuadro y *numeración del cuadro* cuando existen varios de ellos.

### Ejemplo:

Numeración del cuadro



CUADRO No. 1.3

Título



### ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR

Nota preliminar



**Serie Abril 2011 – Abril 2012 (Variaciones Porcentuales)**  
(Base Diciembre 2007=100)

Encabezados



Cuerpo



Meses	INPC
Abril	1,4
Mayo	2,5
Junio	2,5
Julio	2,7
Agosto	2,2
Septiembre	1,6
Octubre	1,8
Noviembre	2,2
Diciembre	1,8
Enero	1,5
Febrero	1,1
Marzo	0,9
Abril	0,8

Fuente



Fuente: INE

**Ejercicio:**

Presentar en un cuadro estadístico la distribución de frecuencias de la variable peso.

**Construcción de gráficos**

Los gráficos facilitan la visualización de las cifras y son ampliamente utilizados en la representación de los datos estadísticos. Cuando se elabora cualquier clase de gráfico se pierde información, pues ya no existen las observaciones originales. Sin embargo, frecuentemente esa pérdida de información es pequeña comparada con la síntesis y facilidad de la interpretación.

Al igual que los cuadros estadísticos, los gráficos deben llevar un *título* que explique de lo que trata la información allí presentada y la *fuentes*. También pueden llevar *notas explicativas*, *numeración correlativa*.

**Algunos tipos de gráficos**

- Diagrama de puntos
- Diagrama de dispersión
- Curvas
- Gráfico de barras
- Gráfico circular
- Pictogramas
- Gráficos especiales para distribuciones de frecuencias de datos cuantitativos:
  - Histograma
  - Diagrama de líneas
  - Polígono de frecuencias
  - Ojiva
  - Diagrama de frecuencias acumuladas
- Diagrama de tallo y hojas
- Diagramas de caja

**Gráfico de barras**

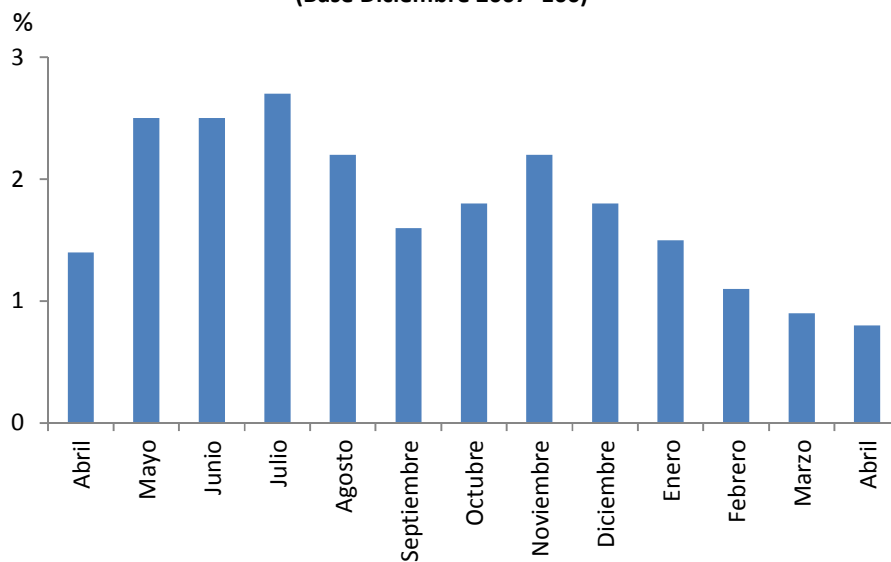
Los gráficos de barras constituyen una herramienta muy adecuada para representar *series cronológicas*, para *datos cualitativos ordinales* y en general para datos donde exista algún orden. En algunas ocasiones también se utiliza en *datos nominales*.

**Construcción del gráfico**

- Paso 1: Establezca un orden (arbitrario cuando la variable es cualitativa nominal) para la colocación, en el eje horizontal de:
  - Las distintas modalidades en el caso de variables cualitativas.
  - El tiempo de ser una serie de tiempo.
- Paso 2: Teniendo en cuenta el valor máximo de la frecuencia (o porcentaje) de los datos, escoja una escala vertical para representar los valores correspondientes.
- Paso 3: En el eje horizontal, para la primera modalidad (o tiempo), dibuje un rectángulo de base cualquiera y altura proporcional al valor de la modalidad.
- Paso 4: Repita el proceso del paso 3 para las demás modalidades.

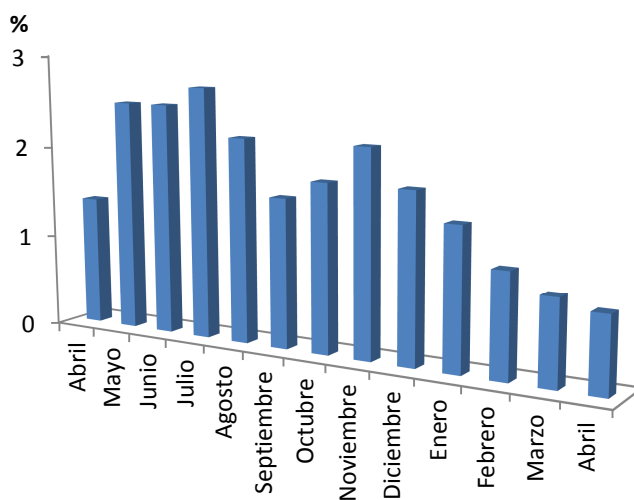
Ejemplo:

**ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR**  
**Serie Abril 2011 – Abril 2012 (Variaciones Porcentuales)**  
**(Base Diciembre 2007=100)**



Fuente: BCV

**ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR**  
**Serie Abril 2011 – Abril 2012 (Variaciones Porcentuales)**  
**(Base Diciembre 2007=100)**

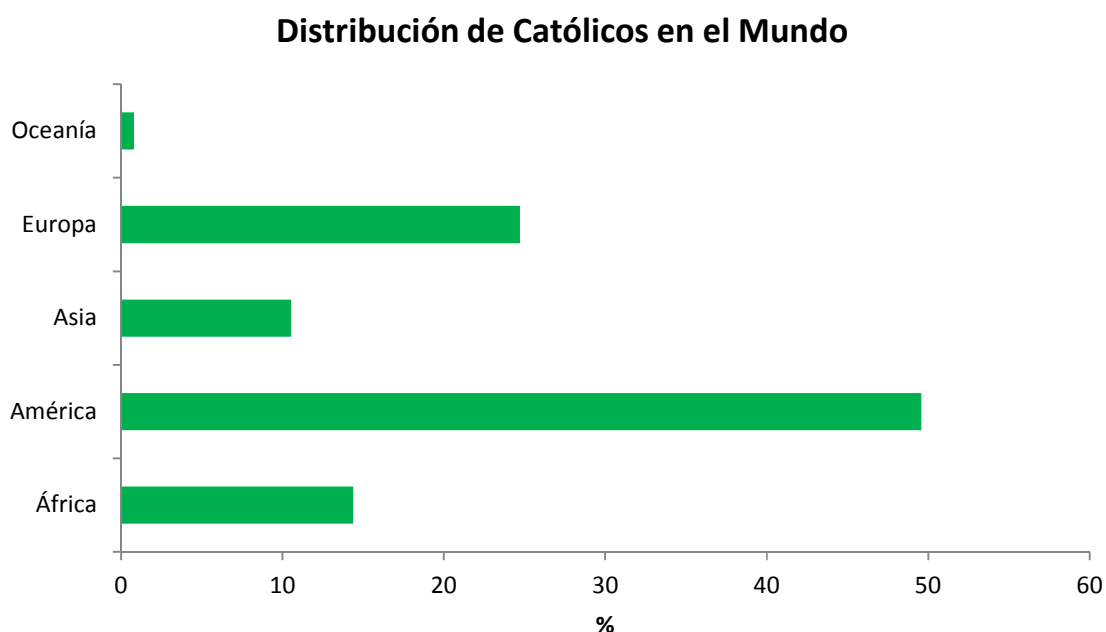


Fuente: BCV

Como se puede apreciar en el gráfico anterior, la presentación puede realizarse en tres dimensiones lo cual puede mejorar la estética del gráfico. Nótese que la información representada en estos gráficos de barras es la misma recogida en la tabla de la página 26. Puede advertirse que resulta más fácil y rápido hacerse una idea del comportamiento del INPC observando el gráfico de barras que mirando los números de la tabla.

El gráfico de barras también puede construirse de tal manera que las barras aparezcan de forma horizontal. Se acostumbra utilizar esta variante cuando se comparan datos cualitativos o datos que se refieren a zonas geográficas.

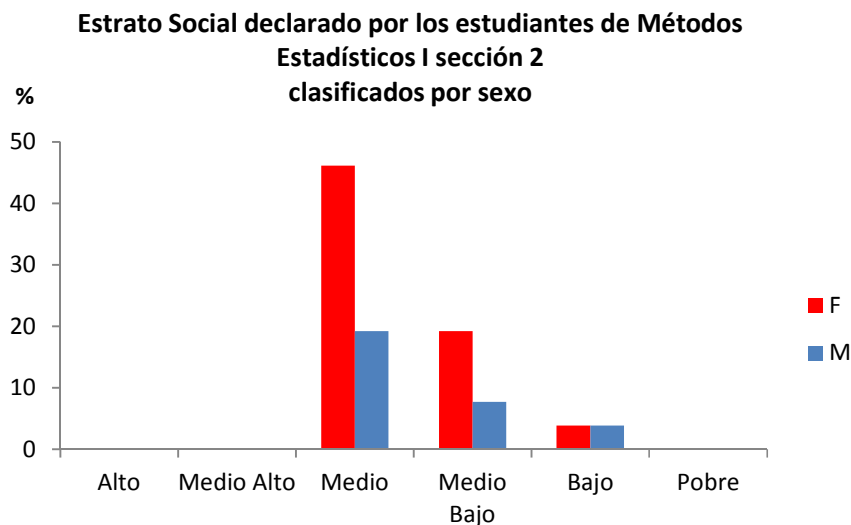
Ejemplo:



Fuente: *Las estadísticas de la Iglesia Católica* (Disponible en: [http://www.vicariadepastoral.org.mx/domund\\_11/imagenes/Estadisticas.pdf](http://www.vicariadepastoral.org.mx/domund_11/imagenes/Estadisticas.pdf) [consulta: 2012, Mayo 7])

Con un gráfico de barras también existe la posibilidad de presentar dos o más variables en un mismo gráfico, de tal manera de que se pueda apreciar el comportamiento individual y además poder hacer comparaciones entre ellas. Veamos esto con un ejemplo:

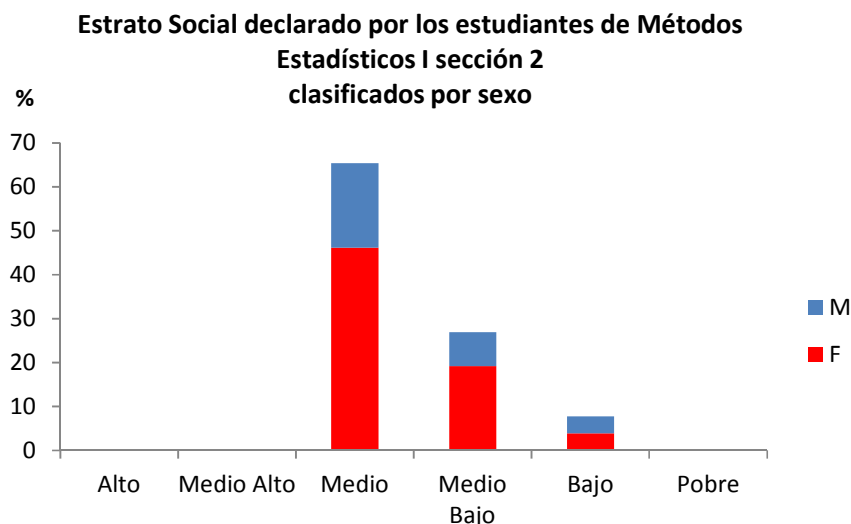
Ejemplo:



*Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas a los estudiantes de Métodos Estadísticos I sección 2 de FACES-ULA. Abril 2012*

Otra variante de esta clase de gráficos es el de *columnas yuxtapuestas o adyacentes*, el cual representa la relación que hay entre los valores o categorías individuales y el total.

Ejemplo:



*Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas a los estudiantes de Métodos Estadísticos I sección 2 de FACES-ULA. Abril 2012*

La desventaja de este tipo de gráfico es que puede ser dificultoso apreciar el comportamiento de la variable de arriba (Masculino en el ejemplo).

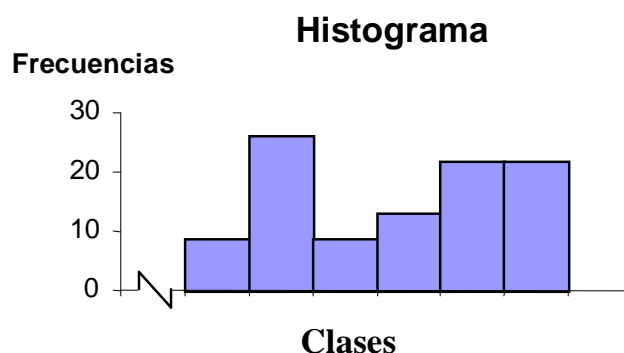
## Histograma

El histograma es el gráfico adecuado para ilustrar el comportamiento de los valores agrupados en intervalos de clase, siendo un gráfico de barras compuesto por varios rectángulos adyacentes, que representan a la tabla de distribución de frecuencias de cierta variable cuantitativa. En el eje horizontal se marcan los intervalos, y cada intervalo es la base de cada rectángulo; en el eje vertical se marcan las alturas de los rectángulos la cual viene dada por las frecuencias respectivas (absolutas simple o relativas)

### Construcción

- Paso 1: En el eje horizontal, marque sucesivamente los límites de cada clase.
- Paso 2: En el eje vertical, marque, en la escala, los valores correspondientes a las frecuencias absolutas o frecuencias relativas de las clases.
- Paso 3: Para la primera clase, construya un rectángulo cuya base es el intervalo de clase y la altura es la frecuencia absoluta simple (o relativa) de esa clase;
- Paso 4: Para la clase siguiente, construya un rectángulo adyacente al primero cuya base es el intervalo de la clase y la altura es la frecuencia absoluta o relativa de esa clase.
- Paso 5: Repita el procedimiento para las demás clases.

En la siguiente figura se representa la apariencia que tendrá un histograma



### *Nota*

Cuando se construyen histogramas, el eje vertical debe mostrar el cero verdadero para no distorsionar o representar equivocadamente el tipo de datos. Sin embargo, no es necesario que el eje horizontal especifique el punto cero del fenómeno de interés. Por razones de estética, el rango de la variable debe constituir la principal porción de la gráfica y, cuando no se incluye el cero, resulta apropiado incluir "fracturas" (↘↗) en el eje.

### Ejercicio:

Construir un histograma para la distribución de frecuencias de la variable peso.

Cuando todas las clases de una distribución de frecuencias tienen la misma amplitud, y el histograma se construye utilizando como altura las frecuencias relativas de clase, se permite la comparación de histogramas correspondientes a distribuciones de frecuencias de datos de la misma naturaleza que difieren en cuanto al número de datos. Cuando se utilizan las frecuencias



absolutas como alturas, no es posible comparar histogramas diferentes excepto en casos especiales.

### Histograma para distribuciones de frecuencias con clases de diferente amplitud

Cuando en la distribución de frecuencias que se ha construido no todas sus clases tienen la misma amplitud, se debe tener en cuenta lo siguiente para la construcción del histograma. La altura de cada rectángulo del histograma debe ser igual a la frecuencia relativa de la clase dividida entre la respectiva amplitud, es decir:

$$h_i = \frac{fr_i}{c_i}$$

De esta manera, las áreas de los rectángulos del histograma serán iguales a la proporción de datos contenidos en la respectiva clase, es decir, a las frecuencias relativas correspondientes. En otras palabras, el tamaño de los rectángulos construidos de este modo, refleja la proporción de datos contenida en esa clase.

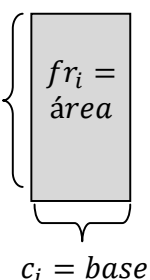
Para entender por qué esto es así, hay que recordar que el área de un rectángulo es:

$$\text{área} = \text{base} \times \text{altura}$$

y si se quiere hallar la altura, despejando queda que:

$$\text{altura} = \frac{\text{área}}{\text{base}}$$

Esta última expresión, en la notación utilizada en distribuciones de frecuencia y para el i-ésimo rectángulo es equivalente a:

$$h_i = \text{altura} \left\{ \begin{array}{l} fr_i = \\ \text{área} \end{array} \right.$$


$$c_i = \text{base}$$

De esta forma, el área total del histograma es igual a 1. Esto es así, ya que al sumar todas las áreas de los rectángulos, lo que realmente se está haciendo es  $\sum fr_i$  que es igual 1.

El histograma construido de este modo es válido tanto para distribuciones de frecuencias cuyas clases tienen todas la misma amplitud como para las que las amplitudes no son todas iguales.

El histograma construido de esta manera permite su comparación con otros histogramas correspondientes a distribuciones de frecuencias de datos de la misma naturaleza que difieran en cuanto al número de datos y a la manera como están establecidas las clases.

**Ejercicio:**

Para la distribución de frecuencias de la variable ingreso:

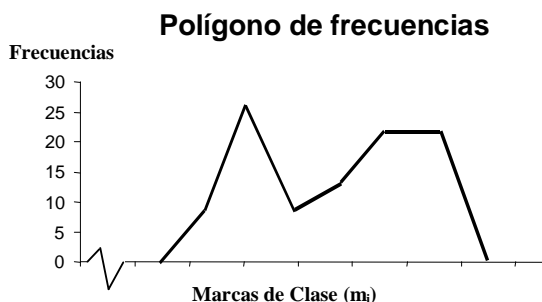
- Construir un histograma con altura de los rectángulos igual a  $fr_i$  (frecuencias relativas)
- Construir un histograma con altura de los rectángulos igual a  $h_i$ .
- ¿Cuál de los dos histogramas anteriores representa adecuadamente a los datos agrupados en la distribución de frecuencias de la variable ingreso? Explique las razones de su elección.

**Nota**

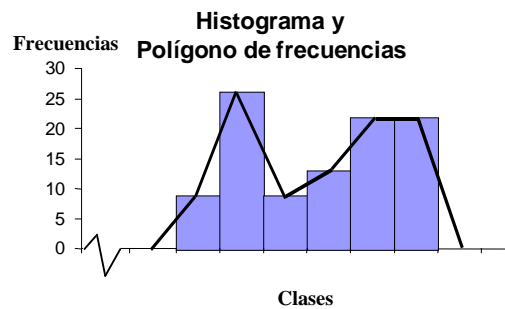
En un histograma, al eliminar los espacios entre las barras se logra que la gráfica lleve consigo una "continuidad" que refleja que los datos son continuos. En los *gráficos de barras* (para datos cualitativos), deben aparecer los espacios entre las barras para evitar que el lector interprete una "continuidad" de los datos, ya que las categorías de las variables cualitativas son ordenadas en general de forma arbitraria y por definición no pueden tomar todos los valores en el eje horizontal.

**Polígono de frecuencias**

Una alternativa para un histograma, es el *polígono de frecuencias*. En el eje horizontal se escriben las *marcas de clase* de cada intervalo y para cada una de estas  $m_i$  se colocan las alturas en el eje vertical, las cuales vienen dadas por las frecuencias respectivas (absolutas simple o relativas). Luego, se marcan los puntos  $(m_i, fr_i$  ó  $m_i, f_i)$  y se une con rectas en el plano cartesiano. Para cerrar la curva resultante con el eje de las abscisas, se crean dos puntos medios ficticios, uno anterior al de la primera clase y otro posterior al de la última clase cada uno con frecuencia igual a cero. De esta manera se obtiene el polígono de frecuencias:



Nótese como el polígono puede obtenerse directamente del histograma:

**Ejercicio:**

Construir un polígono de frecuencias para la distribución de la variable peso.

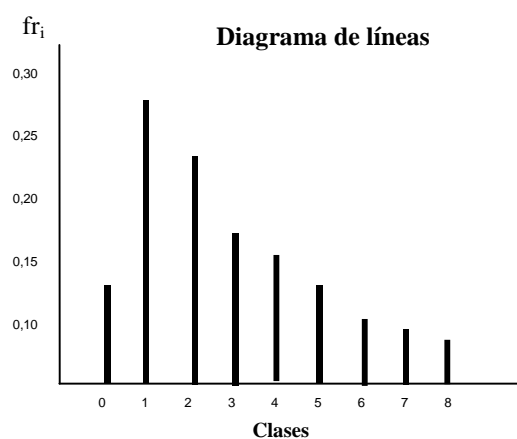
**Diagrama de líneas de frecuencias**

Es el equivalente al histograma en una distribución de frecuencias cuyas clases son valores individuales de la variable.

**Construcción**

- Paso 1: En el eje horizontal, marque sucesivamente las clases.
- Paso 2: En el eje vertical, marque, en la escala, los valores relativos a las frecuencias absolutas o frecuencias relativas de las clases.
- Paso 3: Para la primera clase, trace una línea vertical cuya altura es la frecuencia absoluta simple (o relativa) de esa clase;
- Paso 4: Repita el procedimiento para las demás clases.

En la siguiente figura se representa un diagrama de líneas.

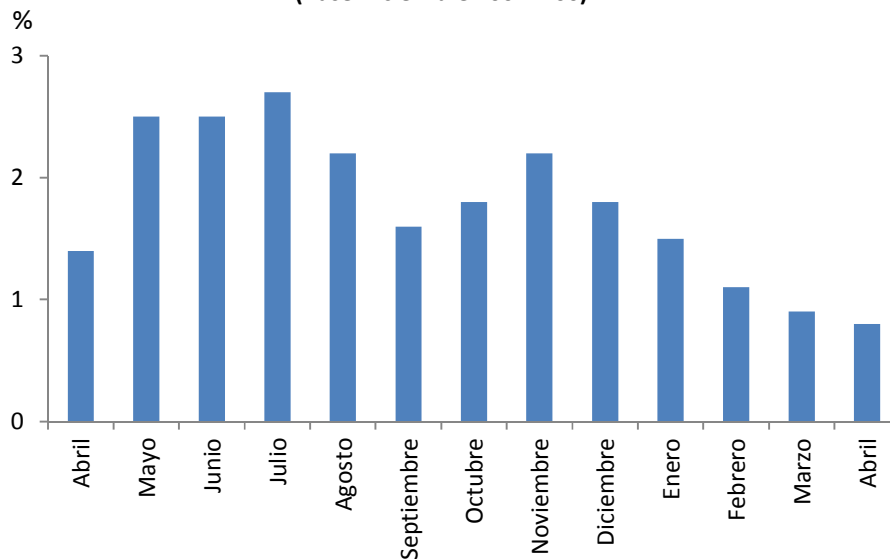
**Ejercicio:**

Construir un diagrama de líneas para la distribución de frecuencias de la variable número de hermanos.

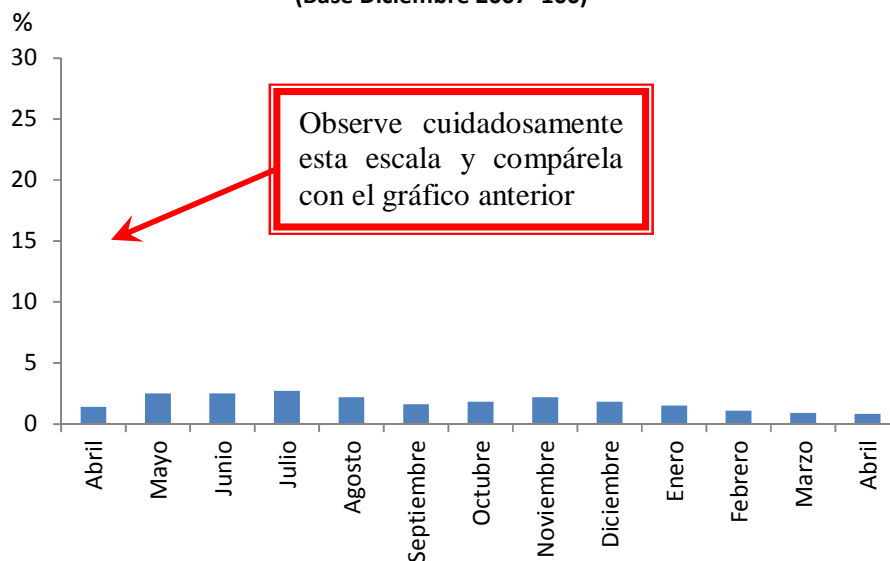
### Gráficos engañosos

¡Cuidado! Cuando se observa un gráfico, particularmente como parte de un anuncio, sea cauteloso. Fíjese en las escalas utilizadas en los ejes vertical y horizontal. Se puede distorsionar la verdad con las técnicas estadísticas, tal como se muestra a continuación:

**ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR**  
**Serie Abril 2011 – Abril 2012 (Variaciones Porcentuales)**  
**(Base Diciembre 2007=100)**



**ÍNDICE NACIONAL DE PRECIOS AL CONSUMIDOR**  
**Serie Abril 2011 – Abril 2012 (Variaciones Porcentuales)**  
**(Base Diciembre 2007=100)**



### Descripción de la forma en que se distribuyen los datos

Los gráficos para distribuciones de frecuencias vistos anteriormente sirven para proporcionar una idea a primera vista acerca de la forma en que se distribuyen los datos. En este sentido se tienen las siguientes definiciones:

#### Distribución simétrica

Una distribución de frecuencias es *simétrica* (o con sesgo cero) con respecto al valor central de la distribución, digamos  $x_0$ , cuando el gráfico a la izquierda de  $x_0$  es el "espejo" de la derecha. En otras palabras, si a la izquierda y a la derecha de  $x_0$  existe la misma cantidad de datos la distribución será simétrica.

#### Distribución asimétrica

Si una distribución no es simétrica, se dice que es asimétrica (o sesgada). Existen dos casos de asimetría:

##### *Asimetría Positiva o por la derecha*

Este tipo de asimetría se presenta cuando existe una mayor concentración de datos en las primeras clases en comparación con las últimas. Se puede visualizar fácilmente cuando el extremo o "cola" de la derecha del gráfico se prolonga más que el de la izquierda.

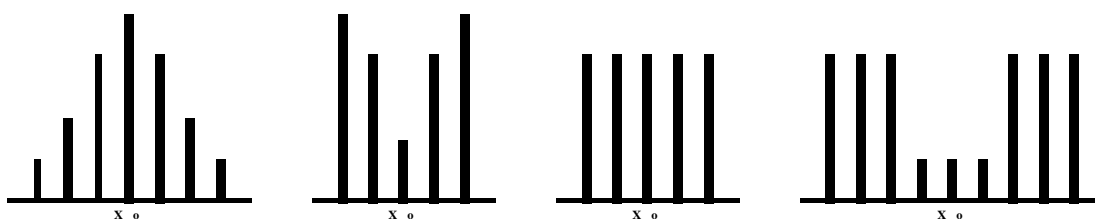
##### *Asimetría Negativa o por la izquierda*

Cuando hay mayor concentración de datos en las últimas clases en comparación con las primeras, es decir, cuando la "cola" izquierda de la curva se prolonga más que la derecha se dice que, la distribución de frecuencias es *asimétrica negativa o por la izquierda*.

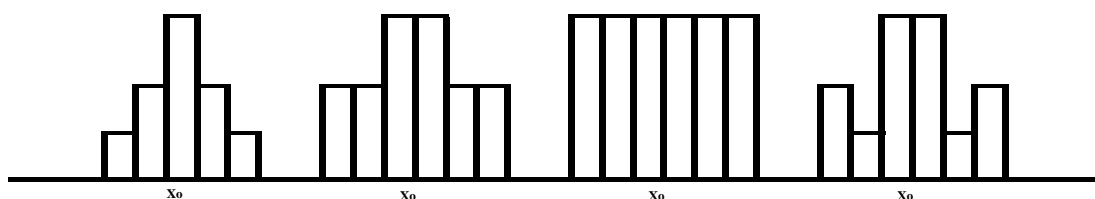
#### Ejemplo:

A continuación se presentan algunos casos de distribuciones simétricas:

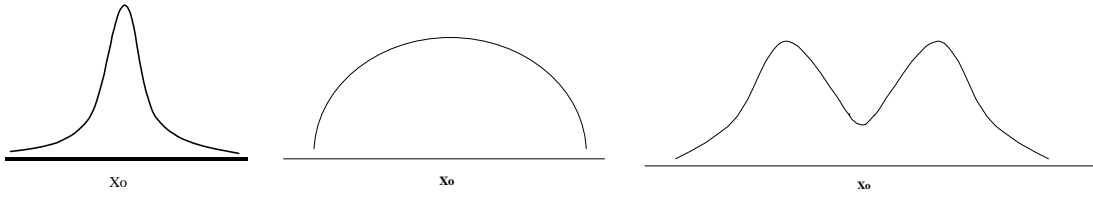
D I A G R A M A S D E L Í N E A S



HISTOGRAMAS

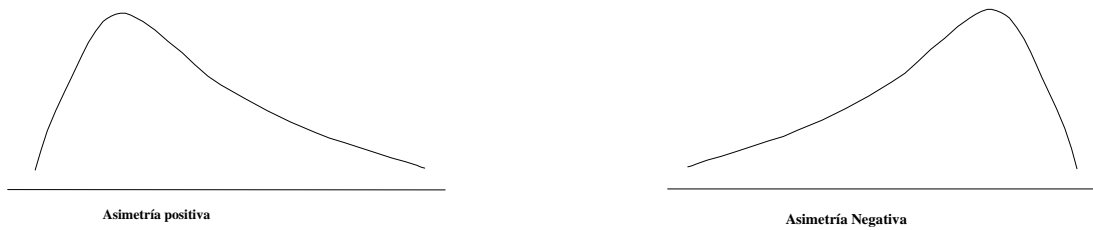


**POLÍGONOS**



**Ejemplo:**

A continuación se presentan dos casos de distribuciones asimétricas:



### Diagrama de Tallo y Hojas<sup>2</sup>

Un diagrama de *tallo y hojas* es una representación visual de los datos que es a la vez una tabla y un gráfico. Es como un histograma horizontal con el cual se puede visualizar rápidamente la distribución de los datos. El diagrama de *tallo y hojas* provee más detalles que un histograma, ya que cada punto del gráfico representa un valor individual de los datos.

#### Construcción

Para ejemplificar la elaboración de un diagrama de tallo y hojas, considérese los datos de la Tabla 1, que representan la duración de 40 baterías de carro similares. Las baterías estaban garantizadas para durar 3 años.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

**Tabla 1. Duraciones de las baterías de un automóvil.**

Primero, se divide cada observación en dos partes que consisten en un tallo y una hoja de tal forma que el primero represente el dígito que es el entero y la hoja corresponda a la parte decimal del número. En otras palabras, para el número 3.7 el dígito 3 se designa como el tallo y el dígito 7 como la hoja. Los cuatro tallos, 1, 2, 3 y 4 quedan listados consecutivamente en el lado izquierdo de la línea vertical de la Tabla 2; las hojas se escriben en el lado derecho de la línea en contraposición al valor de tallo apropiado.

Entonces, la hoja 6 del número 1.6 se escribe a la altura del tallo 1; la hoja 5 del número 2,5 se escribe a la altura del tallo 2; y así sucesivamente. La cantidad de hojas registradas para cada tallo se resume en la columna de frecuencia.

El diagrama de tallo y hojas de la Tabla 2 contiene sólo 4 tallos y, en consecuencia, no proporciona una imagen adecuada de la distribución. Para evitar este problema, se necesita incrementar el número de tallos en la tabla. Una forma sencilla de llevar a cabo esto es escribir

	Tallos	Hojas	Frecuencia
El tallo de 1.9	1	69	2
	2	25696	5
	3	4318514723628297130097145	25
	4	71354172	8

La hoja de 1.9

**Tabla 2. Diagrama de tallo y hojas de las duraciones de las baterías.**

<sup>2</sup> Basado en Walpole, R. y Myers, R. (1993) *Probabilidad y Estadística*. Págs. 59-61.

Tallos	Hojas	Frecuencia
1 S	69	2
2 I	2	1
2 S	5696	4
3 I	431142322130014	15
3 S	8576897975	10
4 I	13412	5
4 S	757	3

**Tabla 3. Diagrama de doble tallo y hojas para las baterías.**

cada valor de tallo dos veces en el lado izquierdo de la línea vertical y entonces registrar las hojas 0, 1, 2, 3 y 4 en el nivel del tallo que les correspondió originalmente; y las hojas 5, 6, 7, 8 y 9 a la altura del tallo registrado la segunda vez. Este diagrama modificado de doble tallo y hojas se muestra en la Tabla 3, donde los tallos que corresponden a las hojas 0, 1, 2, 3 y 4 han sido identificados con el símbolo I (inferior), y los tallos correspondientes a las hojas 5 a 9 por el símbolo S (superior).

Puede lograrse un incremento adicional en el número de tallos al escribir cada valor del tallo cinco veces en el lado izquierdo de la línea vertical donde se podría, ahora, identificar con la letra *a* al tallo para las hojas 0 y 1, con la *b* para las hojas 2 y 3, la *c* para las hojas 4 y 5, la *d* para las 6 y 7 y la *e* para las hojas 8 y 9. Para los datos de la Tabla 1 se usarían entonces, los tallos *1d*, *1e*, *2a*, *2b*, *2c*, *2d* y *2e* para dibujar una tabla de cinco tallos y hojas.

En cualquier problema dado, debe decidirse el valor apropiado de tallos. Esta decisión se toma de manera un poco arbitraria, a pesar de que el tamaño de la muestra sirve de guía. Puede usarse la regla de Sturges como guía de cuantos tallos utilizar, de manera similar a como se hace en la determinación del número de intervalos de clases de una distribución de frecuencias. Usualmente se seleccionan entre 5 y 15 tallos. Entre más pequeña es la cantidad de datos disponibles, menor es el número de tallos seleccionados. Por ejemplo, si los datos consisten en números del 1 al 21 que representan la cantidad de personas en una cafetería en 40 días de trabajo seleccionados aleatoriamente y se decide utilizar un diagrama de doble tallo y hojas, los tallos serían 0I, 0S, 1I, 1S y 2I de tal forma que a la observación más pequeña 1 le corresponde el tallo 0I y la hoja 1, al número 18 le corresponde el tallo 1S y la hoja 8, y a la observación más grande 21 le corresponde el tallo 2I y la hoja 1.

Por otro lado, si los datos consisten en cantidades, desde \$8.800 a \$9.600, las cuales representan los mejores tratos posibles de 100 automóviles nuevos de un cierto distribuidor y se construye un diagrama de tallo y hojas, los tallos serían 88, 89, 90, . . . , 96 y ahora cada hoja tendría dos dígitos. Un vehículo que se vende a \$9,385 tendría un valor de tallo de 93 y la hoja de dos dígitos 85.


Las hojas de múltiples dígitos que pertenecen a un mismo tallo generalmente se separan con comas en el diagrama de tallo y hojas. Los puntos decimales en los datos casi siempre se ignoran cuando todos los dígitos a la derecha del punto decimal representan la hoja. Tal es el caso de la Tabla 2 y la Tabla 3. Sin embargo, si los datos consisten en los números en el rango de 21,8 a



74,9, se podrían seleccionar como tallos los dígitos 2, 3, 4, 5, 6, 7, de tal manera que un número, por ejemplo el 48,3, tendría un valor de tallo de 4 y de hoja 8,3.

También se acostumbra a redondear el valor de la variable a la hoja más cercana para presentarla sólo con un dígito. Por ejemplo, si una variable toma el valor 6,25 tendrá un tallo de 6 y una hoja de 3.

**Nota**

Ante la presencia de valores atípicos el diagrama de tallo y hojas puede requerir una “fractura” (  ) en los tallos.

Ejemplo<sup>3</sup>

A continuación se ilustra el diagrama de tallo y hoja para 25 observaciones del rendimiento por lote de un proceso químico. En el diagrama (a) se han utilizado los números 6, 7, 8, y 9 como tallos. Esto produce muy pocos tallos, con lo que el diagrama no proporciona mucha información sobre los datos. En el diagrama (b) se ha dividido cada tallo en dos partes, con lo que se tiene una presentación más adecuada de los datos. El diagrama (c) ilustra un gráfico en el que cada tallo se ha dividido en cinco partes. En esta gráfica existen muchos tallos, lo que da como resultado una presentación que no dice mucho con respecto a la distribución de los datos.

(a)		(b)		(c)	
Tallo	Hoja	Tallo	Hoja	Tallo	Hoja
6	134556	6I	134	6a	1
7	011357889	6S	556	6b	3
8	1344788	7I	0113	6c	455
9	235	7S	57889	6d	6
		8I	1344	6e	
		8S	788	7a	011
		9I	23	7b	3
		9S	5	7c	5
				7d	7
				7e	889
				8a	1
				8b	3
				8c	44
				8d	7
				8e	88
				9a	
				9b	23
				9c	5
				9d	
				9e	

Las hojas están ordenadas

<sup>3</sup> Tomado de Newbold, P. (1998). *Estadística para los Negocios y la Economía*. Pág. 7.

**Ventajas y Desventajas del Diagrama de Tallo y Hoja**

Los diagramas de tallo y hojas muestran el centro, la dispersión y forma de la distribución de la misma manera en que lo hace un histograma. Sus ventajas incluyen que cada valor de la variable es representado de forma exacta, es muy fácil determinar la mediana y los percentiles; y también es muy fácil de construir con sólo papel y lápiz. Entre las desventajas se encuentra que es difícil comparar distribuciones cuando el número de observaciones en los conjuntos de datos son muy diferentes; y cuando el conjunto de datos es muy grande el diagrama de tallo y hojas se vuelve impráctico.

***Nota***

Observe que en el ejemplo anterior las hojas correspondientes a cada tallo están listadas en orden creciente. El orden de las hojas facilita el uso del diagrama de tallo y hojas en la determinación de la mediana y los percentiles. También se debe notar que todos los posibles valores de los tallos están listados, tengan o no observaciones (hojas) como se ve en el diagrama (c).

**Construcción de una distribución de frecuencias a partir del diagrama de tallo y hojas**

Una distribución de frecuencias en la cual los datos se agrupan en diferentes clases o intervalos, puede construirse con facilidad contando simplemente las hojas que pertenecen a cada tallo y notando que cada uno de ellos define un intervalo. En la Tabla 2 el tallo 1 con dos hojas define el intervalo [1.0-2.0) y contiene dos observaciones; el tallo 2 con 5 hojas define el intervalo [2.0-3.0) y contiene 5 observaciones; el tallo 3 con 25 hojas define el intervalo [3-4) y contiene 25 observaciones y el tallo 4 con 8 hojas define el intervalo [4-5) y contiene 8 observaciones. Para el diagrama de doble tallo y hojas presentado en la Tabla 3, los tallos definen a los 7 intervalos de clase [1.5-2), [2-2.5), [2.5-3), [3-3.5), [3.5-4), [4-4.5) y [4.5-5), con frecuencias 2, 1, 4, 15, 10, 5 y 3, respectivamente.

<b>Clases</b>	$m_i$	$f_i$	$fr_i$
[1.5 – 2.0)	1.75	2	0.050
[2.0 – 2.5)	2.25	1	0.025
[2.5 – 3.0)	2.75	4	0.100
[3.0 – 3.5)	3.25	15	0.375
[3.5 – 4.0)	3.75	10	0.250
[4.0 – 4.5)	4.25	5	0.125
[4.5 – 5.0)	4.75	3	0.075

**Tabla 4. Distribución de frecuencias relativas de las duraciones de las baterías.**

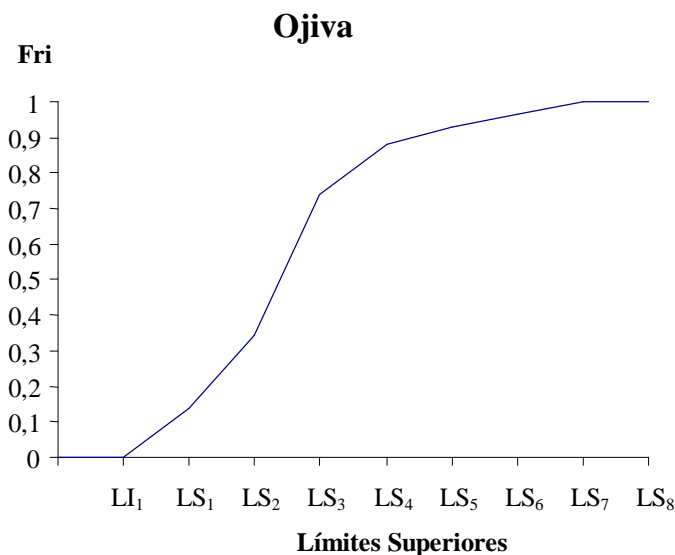
**Ojiva (Polígono de frecuencias acumuladas)**

Este gráfico se emplea en distribuciones de frecuencias cuyas clases son intervalos. Es un tipo especial de gráfico de curvas en el cual se representan las frecuencias acumuladas.

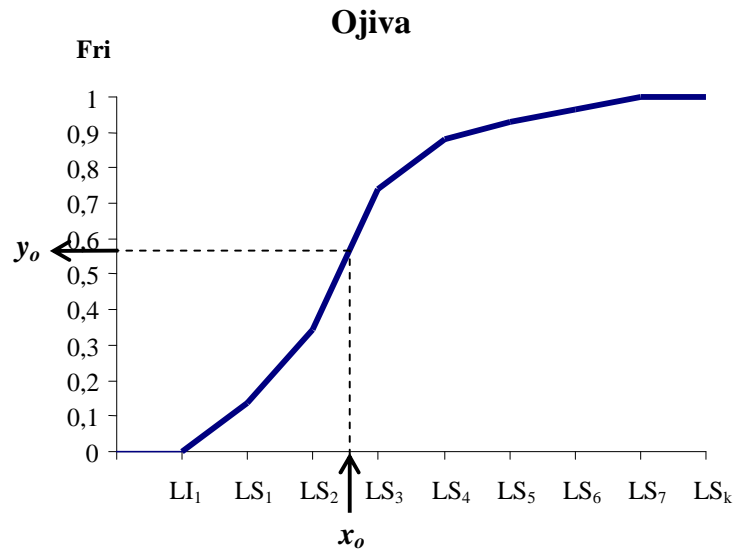
**Construcción**

- Paso 1: En el eje horizontal, marque sucesivamente los límites superiores de cada clase.
- Paso 2: En el eje vertical, marque los valores correspondientes a las frecuencias acumuladas o frecuencias relativas acumuladas.
- Paso 3: Para cada límite superior de clase se marca con un punto su correspondiente frecuencia acumulada.
- Paso 4: El límite inferior de la primera clase también se señala con un punto en el eje horizontal, asignándole una frecuencia acumulada igual a 0.
- Paso 5: Se unen todos los puntos con segmentos de recta.

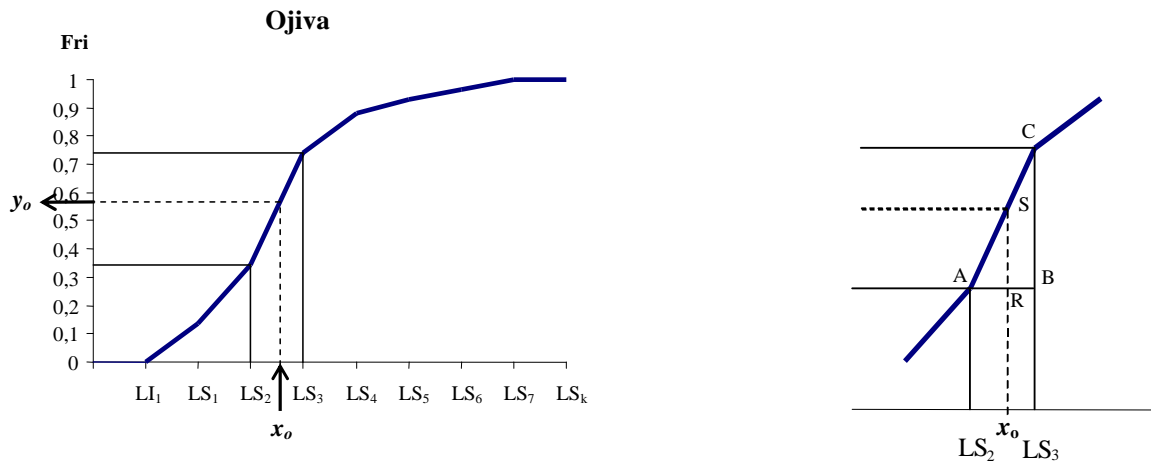
Así se obtiene la *Ojiva*. Nótese que este gráfico es no decreciente.



Las *ojivas* son principalmente usadas para determinar gráficamente y de forma aproximada el número o proporción de datos *que son menores, o que son iguales o mayores* a un valor de interés. Si se usa papel milimetrado para graficar la *ojiva*, se fija el valor  $x_0$  de interés de la variable en estudio el cual es ubicado en el eje horizontal y se levanta desde este valor  $x_0$  una línea perpendicular al eje que llegue hasta la curva. Luego, a partir del punto de intersección se traza una línea paralela al eje de las abscisas; y el punto de corte con el eje vertical,  $y_0$ , representa el número o proporción de datos (dependiendo si la *ojiva* se construyó con las  $F_i$  ó con las  $Fr_i$ ) que son inferiores al valor  $x_0$  especificado.



También se puede encontrar la proporción de datos  $y_0$  que son menores que el valor  $x_0$ , mediante un proceso de interpolación (si no se usa papel milimetrado) usando la propiedad de triángulos semejantes:



Nótese en el gráfico anterior que el triángulo ABC es equivalente con el triángulo ARS, con lo cual se cumple la propiedad:

$$\frac{\overline{AR}}{\overline{AB}} = \frac{\overline{RS}}{\overline{BC}}$$

Que al aplicarse en nuestro caso, tenemos que:

$$\overline{AR} = x_0 - LS_2$$

$$\overline{AB} = LS_3 - LS_2$$

$$\overline{RS} = y_0 - Fr_2$$

$$\overline{BC} = Fr_3 - Fr_2$$

Entonces, sustituyendo en la propiedad de los triángulos semejantes, queda:

$$\frac{x_0 - LS_2}{LS_3 - LS_2} = \frac{y_0 - Fr_2}{Fr_3 - Fr_2}$$

despejando  $y_0$  de la igualdad anterior, se encuentra que:

$$y_0 = \left( \frac{x_0 - LS_2}{LS_3 - LS_2} \right) * (Fr_3 - Fr_2) + Fr_2$$

Así, a través de ese método de interpolación se puede encontrar una aproximación a la proporción de datos, igual a  $y_0$ , que es menor que el valor  $x_0$ .

### Ejercicios:

Usando la distribución de frecuencias construida en clase para la variable peso, encuentre mediante el método gráfico de interpolación:

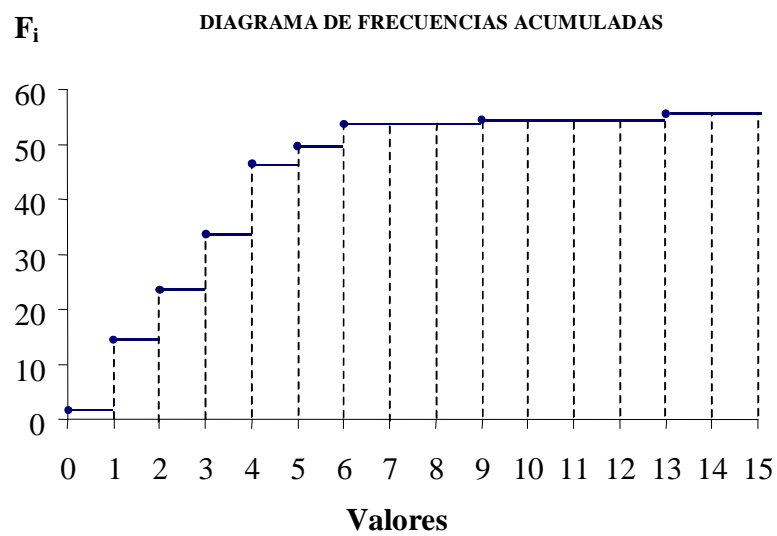
1. La proporción de estudiantes que pesan menos de 78 Kg.
2. La proporción de estudiantes cuyos pesos son menores a 56 Kg.
3. La proporción de estudiantes que tienen un peso mayor o igual a 56 Kg.
4. El porcentaje de estudiantes que pesan menos de 56 Kg.
5. El valor del peso por encima del cual se encuentra el 50% de los pesos de los estudiantes.
6. El peso tal que el 10% de los estudiantes está por debajo de él.
7. El peso tal que la mitad de los datos está por debajo de su valor.

### **Diagrama de Frecuencias Acumuladas (Gráfico de escalera)**

El gráfico equivalente a la *ojiva* en el caso de *distribuciones de frecuencias cuyas clases son valores individuales de la variable en estudio* se denomina *diagrama de frecuencias acumuladas*.

### Construcción

- Paso 1: En el eje horizontal, marque sucesivamente los valores de la variable que representan las clases.
- Paso 2: En el eje vertical, marque los valores correspondientes a las frecuencias acumuladas o frecuencias relativas acumuladas (o porcentaje).
- Paso 3: A cada valor de la variable se le representa su frecuencia acumulada mediante una línea horizontal que se prolonga hasta donde está señalado el próximo valor de la variable.
- Paso 4: Al trazar las líneas anteriores, se les coloca un punto al comienzo. Esto indica que al correspondiente valor en el eje horizontal le corresponde esa frecuencia acumulada.



Nótese que el gráfico suministra visualmente el número de datos *menores o iguales* que un valor particular de la variable en estudio.

## Comentarios finales

En las aplicaciones prácticas, el objetivo de construir una distribución de frecuencias y su respectivo histograma es conseguir información relevante sobre los datos. En este sentido, la decisión más difícil, pero inevitable, es cuanto ha de detallarse. Si se realiza una presentación muy poco detallada, es decir con muy pocas clases, se pueden ocultar características importantes, mientras que si se cae en el otro extremo, podríamos perdernos en un exceso de detalle. La mejor guía a seguir es el sentido común, aunque pueden enumerarse unas cuantas reglas generales:

1. Como se había comentado antes, para lograr una interpretación más fácil es preferible establecer intervalos de igual longitud. Sin embargo, en algunas ocasiones habrá que descartar este principio. Si un conjunto de datos tiene muchas observaciones contenidas en muy pocos intervalos de clase, mientras que las otras están muy dispersas en el resto de las clases, será preferible dividir en intervalos de longitud pequeña la zona donde las observaciones están más concentradas, y en intervalos más amplios las observaciones fuera de esta zona. Si se hace esto, es importante tener muy en cuenta que son las *áreas* y no las alturas de los rectángulos del histograma, las que han de ser proporcionales a las frecuencias.

2. Es importante asegurarse que los puntos medios de los intervalos sean representativos de los miembros de esa clase. Por ejemplo, muchos artículos en las tiendas tienen precios de Bs. 9.999, Bs. 10.999, etc. Si se clasifican los precios en intervalos [Bs. 9.000 - Bs. 10.000), [Bs. 10.000 - Bs. 11.000), etc., es muy probable que en cada intervalo de clase predominen los precios próximos al límite superior. Una mejor solución consistiría en establecer clases como: [Bs. 9.500 - Bs. 10.500), [Bs. 10.500 - Bs. 11.500) y así sucesivamente.

Una razón para elegir puntos medios de clase que sean representativos de los valores de los miembros de esa clase, es que el histograma tendrá un aspecto visual más fidedigno. Además como se verá más adelante, en muchos casos se calculan medidas de centralización y dispersión para datos agrupados. Estos cálculos dependen del supuesto de que *los puntos medios de cada intervalo de clase son representativos de la clase*.

3. Muchas veces, la decisión más difícil de tomar es decidir el número de clases a incluir. Si el número de clases es demasiado pequeño, la clasificación resultante puede esconder aspectos importantes de los datos. Si hay demasiadas clases, puede resultar un gráfico quebrado y desigual, difícil de interpretar. En general, como se había recomendado antes, debe usarse un número de clases mayor que cinco y menor que 15. Para conjuntos de datos muy grandes, con muchas observaciones, será razonable establecer más clases. Subdividir, por ejemplo, un conjunto de 20 observaciones en 15 clases pequeñas conllevaría a tener muchas clases vacías o casi vacías. Esto puede ser un problema menos grave si se tienen 200 observaciones.

Incluso si se tienen en cuenta los factores enunciados no siempre estará clara la elección del número de intervalos. En muchos casos, una buena idea es probar varias posibilidades y ver cuál de los histogramas resultantes presenta un aspecto más claro.