Probabilidad y estadística

Docente a Cargo:
Ing. Pablo Vega
Mail: pablorvega@gmail.com

- Materiales de estudio: diapositivas de clases y cualquier texto de Estadística básica.

  Walpole, R. E., Myers, R. H., Myers, S. L. y Ye, K. (2012) Probabilidad y Estadística para Ingeniería y Ciencias. Novena edición. Pearson Educación, México.

- Software estadístico: MINITAB 18 o Excel

- Evaluación: examen escrito individual

UNIDADES TEMATICAS

1. Metodología estadística
2. Medias de posición y dispersión
3. Algebra de probabilidades
4. Variable aleatoria
5. Modelos especiales de probabilidad
6. Elementos de muestreo
7. Teoría de la estimación estadística
8. Contraste, prueba o docimasia de hipótesis

UNIDADES TEMATICAS

1. Metodología estadística
2. Medias de posición y dispersión
3. Algebra de probabilidades
4. Variable aleatoria
5. Modelos especiales de probabilidad
6. Elementos de muestreo
7. Teoría de la estimación estadística
8. Contraste, prueba o docimasia de hipótesis

Organización y presentación de datos estadísticos

- **Distribución unidimensional:** cuando de las unidades estadísticas de una población bajo estudio se considera una sola variable, se dice que se esta en presencia de una investigación unidimensional. Si son dos las variables, bidimensional y así sucesivamente.

- **Serie simple:** se denomina de esta forma al conjunto de datos obtenidos en bruto luego de realizado el relevamiento pertinente. Se denota con $\mathbf{x_i}$ a cada observación a los efectos de su individualización a través del subíndice **i**, variando desde **1** hasta **n**, siendo **n** el total de las observaciones. En forma simbólica para **n** elementos:

$$x_1, x_2, x_3, \ldots\ldots, x_n \Rightarrow x_i \ / \ i=1,n$$

Ejemplo: de los 14 empleados de una empresa se quiere saber el numero de hijos que cada uno de ellos tiene:

$$x_1=3, x_2=2, x_3=3, x_4=2, x_5=1, x_6=0, x_7=5, x_8=2, x_9=3, x_{10}=3, x_{11}=1, x_{12}=0, x_{13}=1, x_{14}=3$$

- **Distribución de frecuencias:** podemos encontrar dos situaciones de acuerdo a la variable con la que estemos trabajando.

1. Variable discreta
2. Variable continua

1. **Distribución de frecuencia de variable discreta:** si los datos de una serie simple se corresponden con un variable discreta y sus valores son homogéneos, podemos definir algunas variables:

y_i: valores iguales observados, variando el subíndice i desde 1 hasta k, siendo k el numero de valores distintos de la serie simple.

n_i: numero de veces que determinado valor aparece repetido en la serie simple, denominada **frecuencia absoluta** que la denotaremos n_i. Tiene la siguientes propiedades:

a) Son números enteros comprendidos entre **0** y **n**, siendo **n** el total de las observaciones.
b) $\sum n_i = n$ (numero de observaciones)

h_i (frecuencia relativas): se obtiene del cociente entre las frecuencias absolutas asociadas a cada valor de las variables, y el total de observaciones:

$$h_i = n_i / n$$

Indican la importancia relativa de cada valor de la variable en relación al total de observaciones, pudiéndose interpretar en términos porcentuales. Verifican las siguientes propiedades:

a) Son números fraccionarios comprendidos en el intervalo: $0 \leq h_i \leq 1$
b) $\sum h_i = 1$

H_i (frecuencia relativa acumulada): es representativa del total acumulado de frecuencia relativa

$$H_k = 1$$

N_i (frecuencia absolutas acumulada): es representativa del total acumulado de frecuencia absolutas

N_k = n (total de observaciones)

Expresión genérica de una distribución de frecuencias de variable discreta

Suponiendo **k** valores distintos de la variable observada en la serie simple, de forma general una distribución de frecuencias de variable discreta responde a la siguiente estructura.

y_i	n_i	h_i	N_i	H_i
y_1	n_1	h_1	N_1	H_1
y_2	n_2	h_2	N_2	H_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
y_k	n_k	h_k	N_k	H_k
$\sum$	n	1		

$x_1=3, x_2=2, x_3=3, x_4=2, x_5=1, x_6=0, x_7=5, x_8=2, x_9=3, x_{10}=3, x_{11}=1, x_{12}=0, x_{13}=1, x_{14}=3$

N° de hijos (y_i)	N° de empleados(n_i)	h_i	N_i	H_i
0	2	0,15	2	0,15
1	3	0,21	5	0,36
2	3	0,21	8	0,57
3	5	0,36	13	0,93
4	0	0	13	0,93
5	1	0,07	14	1
∑	14	1		

$x_1=3$, $x_2=2$, $x_3=3$, $x_4=2$, $x_5=1$, $x_6=0$, $x_7=5$, $x_8=2$, $x_9=3$, $x_{10}=3$, $x_{11}=1$, $x_{12}=0$, $x_{13}=1$, $x_{14}=3$

2. **Distribución de frecuencia de variable continua:** cuando los datos de una serie simple se corresponden con un variable discreta con valores heterogéneos, o bien con una variable continua, los valores de la variable se expresan en grupos denominados genéricamente “intervalos de clase”, en donde los extremos izquierdo y derecho de la forma y_{i-1} e y_i respectivamente.

Los intervalos de clase no necesitan ser del mismo tamaño, pero para facilitar la interpretación grafica de la distribución de frecuencias, así como para el calculo de las distintas medidas estadísticas, es conveniente que sean del mismo tamaño. Cabe destacar que los intervalos se consideran “semiabiertos por la derecha”.

Los conceptos de frecuencias absolutas y relativas simples se corresponden con el numero de observaciones asociadas a los valores de la variable de cada intervalo. En el caso de las acumuladas, son representativas del total acumulado de frecuencias (absolutas o relativas), hasta el valor de la variable correspondiente al extremo derecho del intervalo donde se encuentra, **sin considerar el valor de dicho extremo,** ya que son semiabiertos por la derecha.

Intervalo semiabierto a derecha (o semicerrado a izquierda)

Es el conjunto de números reales formado por a y los números comprendidos entre a y b.

$[a, b) = \{ x / a \leq x < b\}$

Para la construcción de una distribución de frecuencias de variables continua se procede de la siguiente forma:

a) **Recorrido:** se calcula el **recorrido** de la variable **(R)**, dado por la diferencia entre el valor mayor y el valor menor de los observados de la variable en la serie simple. Ejemplo, se el mayor valor es 93 y el menor es 43.

$$R = 93 - 43 = 50$$

b) Si se tiene como **dato el numero de intervalos**, es necesario determinar la **amplitud de los mismos (c_i)**, de la forma:

$$\text{Amplitud } (c_i) = \text{Recorrido} / \text{N° de intervalos}$$

Cabe destacar que la amplitud de un intervalo esta dado por la diferencia entre el extremo derecho e izquierdo de cada intervalo de clase, es decir:

$$c_i = y_i - y_{i-1}$$

Así por ejemplo, si se desea una distribución de frecuencia con 5 intervalos de clase, considerando el Recorrido anterior, se tendrá:

$$\text{Amplitud} = 50 / 5 = 10$$

Es conveniente que la amplitud sea un numero divisible por 2, en caso de no serlo, debe ampliarse el recorrido

b) Si se tiene como **dato la amplitud de los intervalos**, debe determinarse **el numero de los mismos**, de la forma:

$$\text{N° de intervalos} = \text{Recorrido} / \text{Amplitud}$$

Así por ejemplo, considerando el Recorrido anterior y una amplitud de los intervalos de 10 se tendrá:

$$\textbf{N° de intervalos} = 50 / 10 = 5$$

En el caso de que al efectuar el cociente se obtenga un numero decimal, debe ampliarse el Recorrido, para de esta manera lograr un numero entero.

c) La **marca de clase** (y_i) esta dada por el punto medio de cada intervalo de clase y se calcula a los efectos de determinar un valor representativo de la variable de cada intervalo de clase, para obtener las distintas medidas estadísticas. Su valor surge del cociente:

$$y_i = (y_{i-1} + y_i) / 2$$

Expresión genérica de una distribución de frecuencias de variable continua

Suponiendo una distribución de frecuencia de variable continua con **k** intervalos de clases, simbólicamente la podemos expresar de la forma:

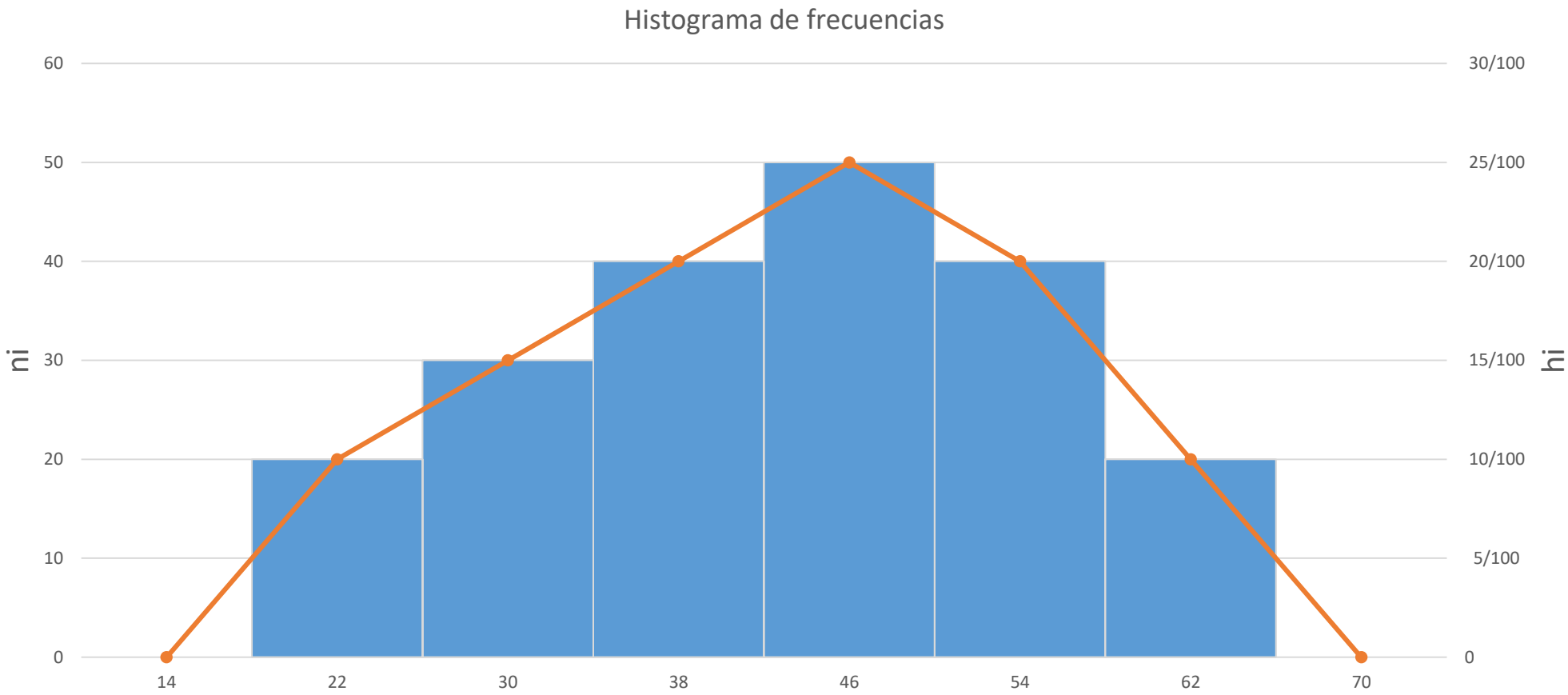
Inter. de clase		Marca de clase	Frec. Abs.	Frec. Rel.	Frec. Abs. Ac.	Frec. Rel. Ac.
y'_{i-1}	y'_i	y_i	n_i	h_i	N_i	H_i
y'_0	y'_1	y_1	n_1	h_1	N_1	H_1
y'_1	y'_2	y_2	n_2	h_2	N_2	H_2
-	-	-	-	-	-	-
y'_{k-1}	y'_k	y_k	n_k	h_k	N_k	H_k
		$\sum$	n	1		

Ejemplo: se clasifico al personal de una empresa de acuerdo a sus edades, obteniéndose la siguiente **Distribución de frecuencias de variable continua.**

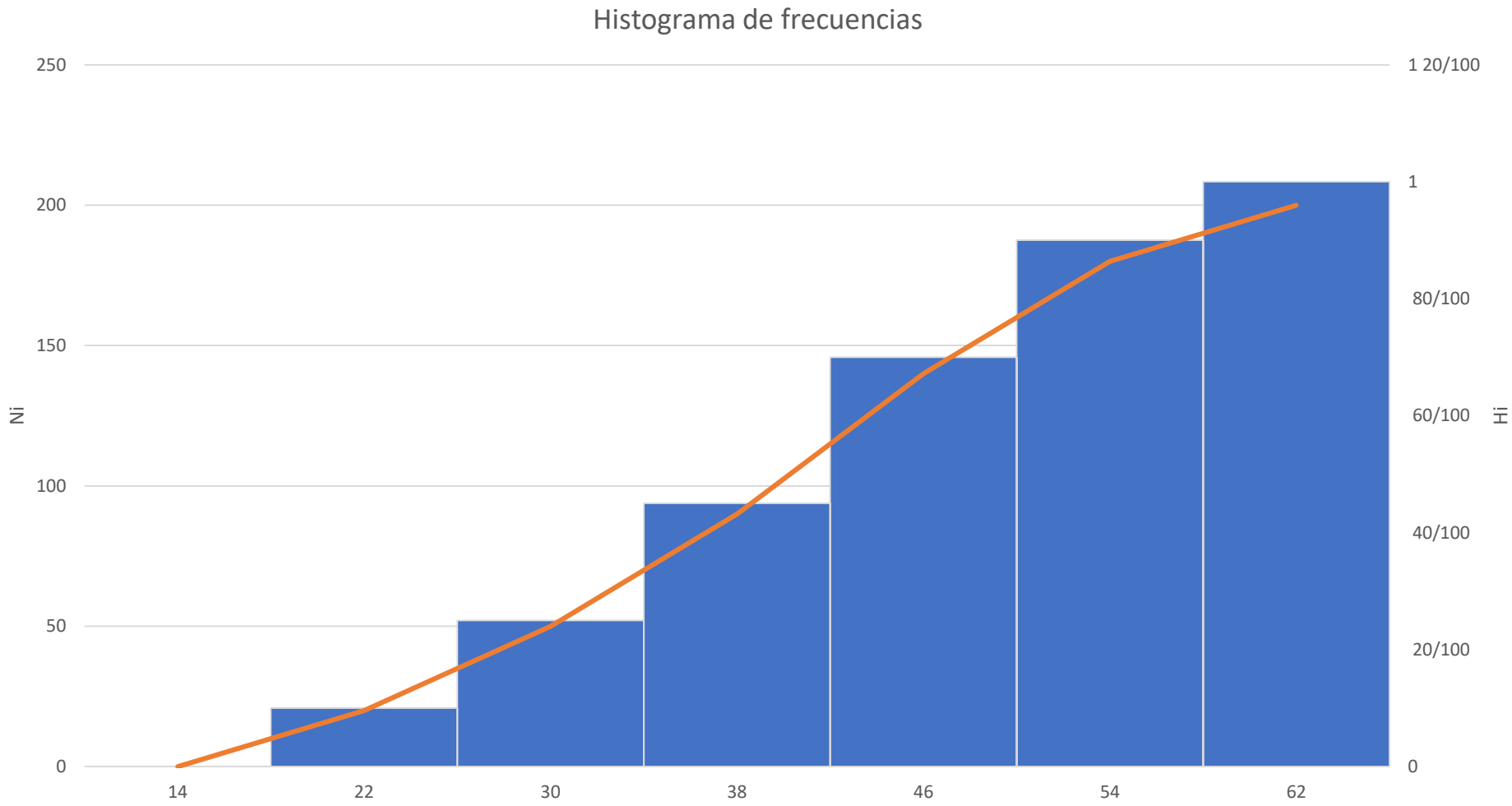
Inter. de clase		Marca de clase	n° de empleados	Frec. Rel.	Frec. Abs. Ac.	Frec. Rel. Ac.
y'_{i-1}	y'_i	y_i	n_i	h_i	N_i	H_i
18	26		20			
26	34		30			
34	42		40			
42	50		50			
50	58		40			
58	66		20			
			200			

Ejemplo: se clasifico al personal de una empresa de acuerdo a sus edades, obteniéndose la siguiente **Distribución de frecuencias de variable continua.**

Inter. de clase		Marca de clase	n° de empleados	Frec. Rel.	Frec. Abs. Ac.	Frec. Rel. Ac.
y'_{i-1}	y'_i	y_i	n_i	h_i	N_i	H_i
		14	0	0	0	
18	26	22	20	0,1	20	0,1
26	34	30	30	0,15	50	0,25
34	42	38	40	0,2	90	0,45
42	50	46	50	0,25	140	0,7
50	58	54	40	0,2	180	0,9
58	66	62	20	0,1	200	1
		70	0	0		
			200	1		



Uniendo los puntos medios de cada rectángulo del histograma, que se corresponden con los valores de las marcas de clase de cada intervalo, queda determinada la poligonal de frecuencias absolutas o relativas simples la que sirve para caracterizar gráficamente a una distribución de frecuencias de variable continua, siendo el área entre la poligonal y el eje de las abscisas igual a la suma de las superficies de los rectángulos del histograma, que da lugar a la “densidad de frecuencias”.



La poligonal de frecuencias acumuladas permite realizar interpolaciones lineales graficas

Actividad 1:

Las velocidades de 55 automóviles fueron medidas con un radar en una calle de cierta ciudad:

1. Organice los datos en una tabla de frecuencias.
2. Grafique adecuadamente los datos.

27	23	22	38	43	24	35	26	28	18	20
25	23	22	52	31	30	41	45	29	27	43
29	28	27	25	29	28	24	37	28	29	18
26	33	25	27	25	34	32	36	22	32	33
21	23	24	18	48	23	16	38	26	21	23

Actividad 2:

Se registró el nivel de absorción de agua de una muestra de 20 pedazos de fibra de algodón. Los valores son los siguientes:

1. Identifique la variable y clasifíquela.
2. Construya una tabla de frecuencias. Grafique.

21	19	22	19	18	20	23	19	19	20
19	20	21	22	21	20	22	20	21	20

- **Representación grafica y tabular:**

Los gráficos y las tablas o cuadros estadísticos permiten presentar a un conjunto de datos en forma ordenada y resumida, de forma tal que una mirada rápida hace posible extraer alguna conclusión acerca del comportamiento de las variables estudiadas.

Vendedor	Sector	Numero de clientes	Total de ventas ($)
Pérez	Electrodomésticos	450	420000
García	Indumentaria	1200	470000
López	Perfumería	750	212500
Campana	Electrodomésticos	560	680000
Rodriguez	Decoración	480	582000
Gomez	Decoración	550	660000
Arancibia	Perfumeria	730	175000
Guitierrez	Indumentaria	1350	585000
Martinez	Electrodomesticos	590	730000
Gonzalez	Indumentaria	1430	720000
Constancio	Decoracion	620	715000
Arias	Perfumeria	685	195000

1) Ordenar los datos en forma ascendente considerando el **Total de Ventas**

2) Ordenar los datos en forma descendente considerando el **numero de clientes**

3) Ordenar los datos de los **vendedores** alfabéticamente.

4) Ordenar los datos por **sector** alfabéticamente.

1) Ordenar los datos en forma ascendente considerando el **Total de Ventas**

Vendedor	Sector	Numero de clientes	Total de ventas ($)
Arancibia	Perfumeria	730	175000
Arias	Perfumeria	685	195000
López	Perfumería	750	212500
Perez	Electrodomésticos	450	420000
Garcia	Indumentaria	1200	470000
Rodriguez	Decoración	480	582000
Guitierrez	Indumentaria	1350	585000
Gomez	Decoracion	550	660000
Campana	Electrodomesticos	560	680000
Constancio	Decoracion	620	715000
Gonzalez	Indumentaria	1430	720000
Martinez	Electrodomesticos	590	730000

2) Ordenar los datos en forma descendente considerando el **numero de clientes**

Vendedor	Sector	Numero de clientes	Total de ventas ($)
González	Indumentaria	1430	720000
Guitierrez	Indumentaria	1350	585000
Garcia	Indumentaria	1200	470000
Lopez	Perfumería	750	212500
Arancibia	Perfumería	730	175000
Arias	Perfumería	685	195000
Constancio	Decoracion	620	715000
Martinez	Electrodomesticos	590	730000
Campana	Electrodomesticos	560	680000
Gomez	Decoracion	550	660000
Rodriguez	Decoracion	480	582000
Perez	Electrodomesticos	450	420000

3) Ordenar los datos de los **vendedores** alfabéticamente.

Vendedor	Sector	Numero de clientes	Total de ventas ($)
Arancibia	Perfumeria	730	175000
Arias	Perfumería	685	195000
Campana	Electrodomésticos	560	680000
Constancio	Decoracion	620	715000
Garcia	Indumentaria	1200	470000
Gomez	Decoracion	550	660000
Gonzalez	Indumentaria	1430	720000
Guitierrez	Indumentaria	1350	585000
Lopez	Perfumeria	750	212500
Martinez	Electrodomesticos	590	730000
Perez	Electrodomesticos	450	420000
Rodriguez	Decoracion	480	582000

4) Ordenar los datos por **sector** alfabéticamente.

Vendedor	Sector	Numero de clientes	Total de ventas ($)
Constancio	Decoracion	620	715000
Gómez	Decoracion	550	660000
Rodriguez	Decoracion	480	582000
Campana	Electrodomesticos	560	680000
Martinez	Electrodomésticos	590	730000
Perez	Electrodomesticos	450	420000
Garcia	Indumentaria	1200	470000
Gonzalez	Indumentaria	1430	720000
Guitierrez	Indumentaria	1350	585000
Arancibia	Perfumeria	730	175000
Arias	Perfumeria	685	195000
Lopez	Perfumeria	750	212500

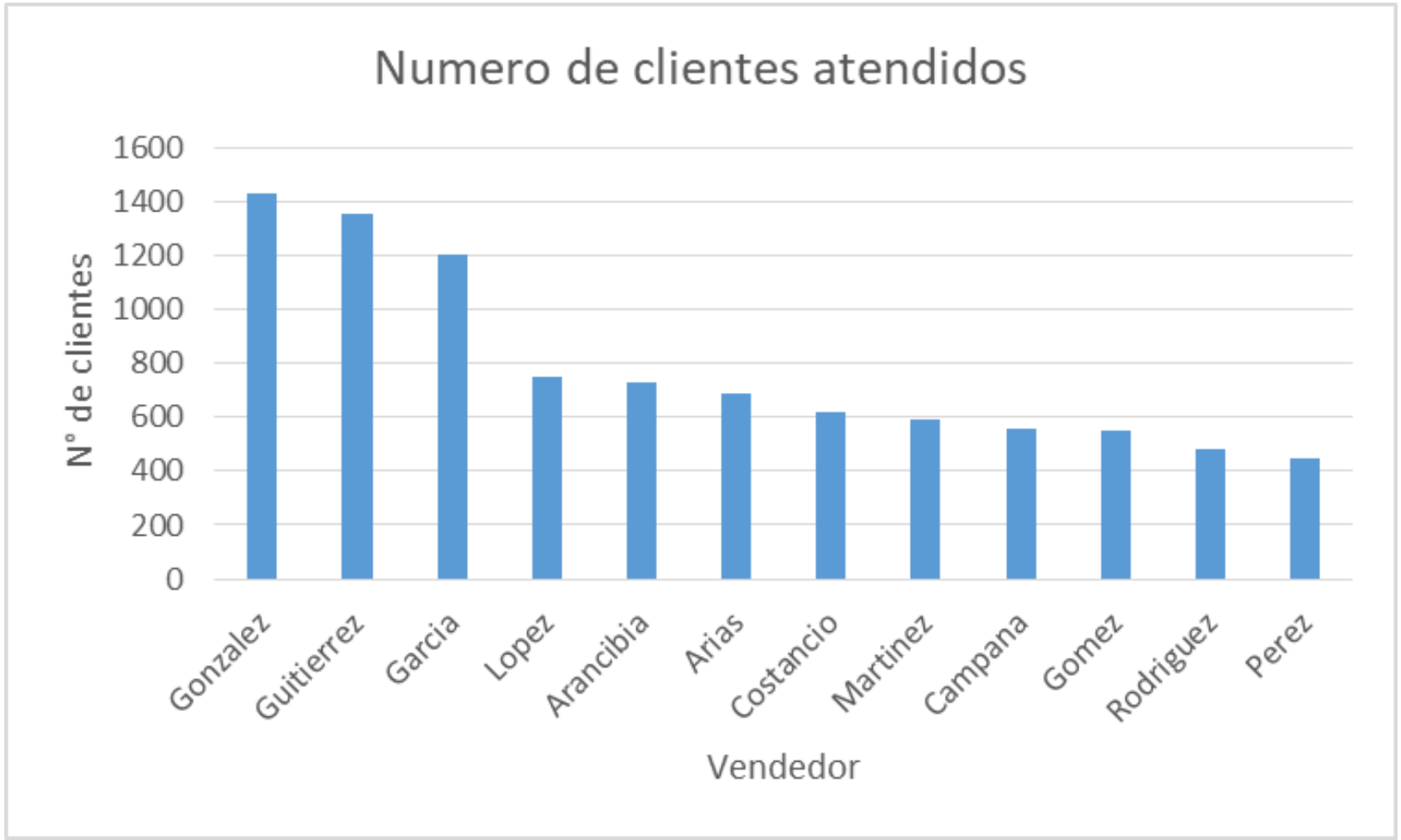
- **Representaciones graficas**

De acuerdo a la naturaleza de los datos tenemos distintas alternativas en relación a los gráficos, debiendo escogerse aquella alternativa mas adecuada para cada caso. Al respecto siempre debe tenerse en consideración los siguientes aspectos:

1) Todo grafico debe tener titulo el que debe tender a contextualizar el grafico facilitando su lectura, debiendo ser claro permitiendo explicitar las variables que representan, unidad de medida de las variables, periodos, etc.
2) Todo grafico debe tener titulo en los ejes para facilitar su interpretación.

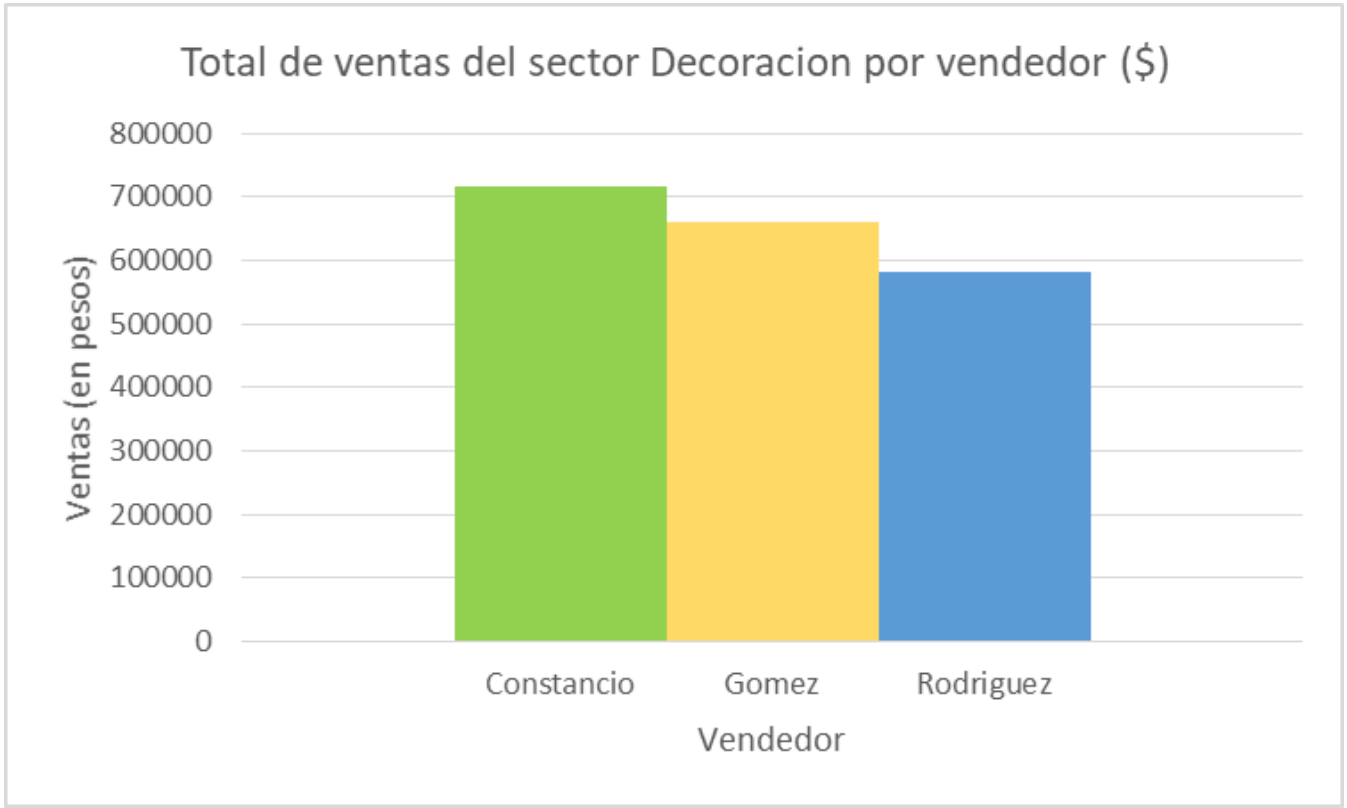
1) Ordenar los datos en forma ascendente considerando el **Total de Ventas**

2) Ordenar los datos en forma descendente considerando el **numero de clientes**



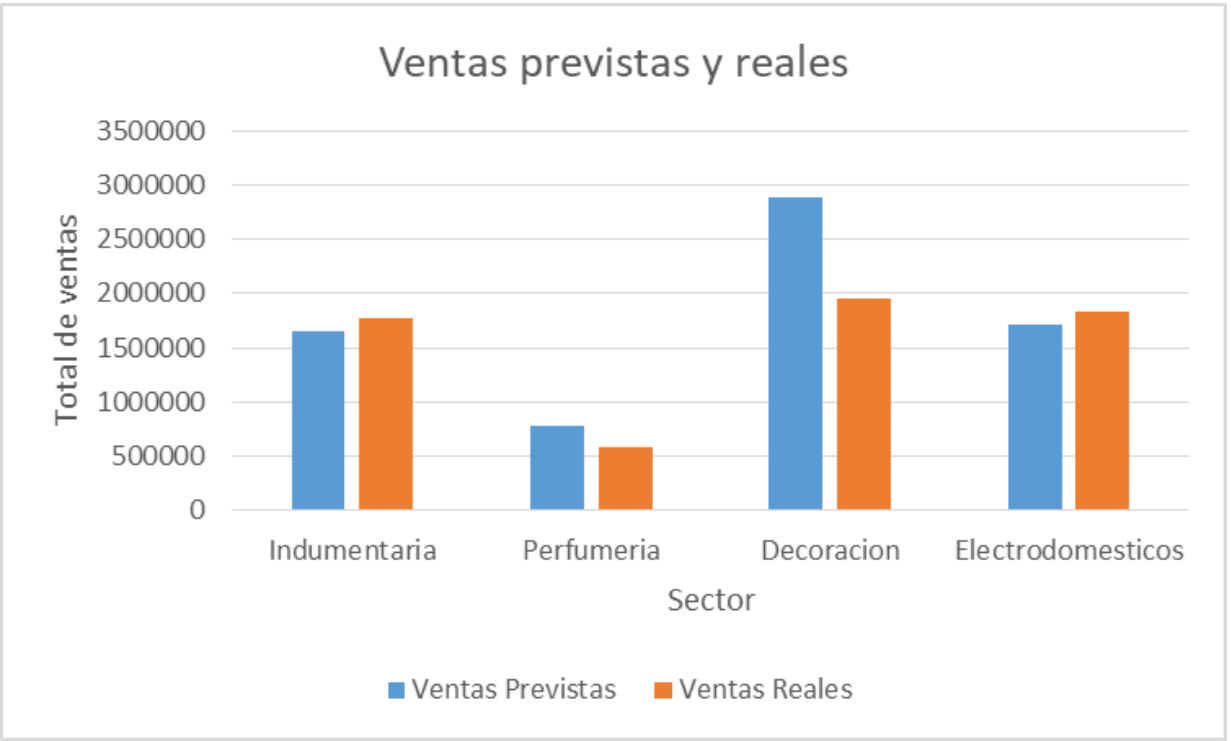
3) Considerando los datos de la siguiente tabla, correspondiente a las ventas del **sector Decoración** por **Vendedor**

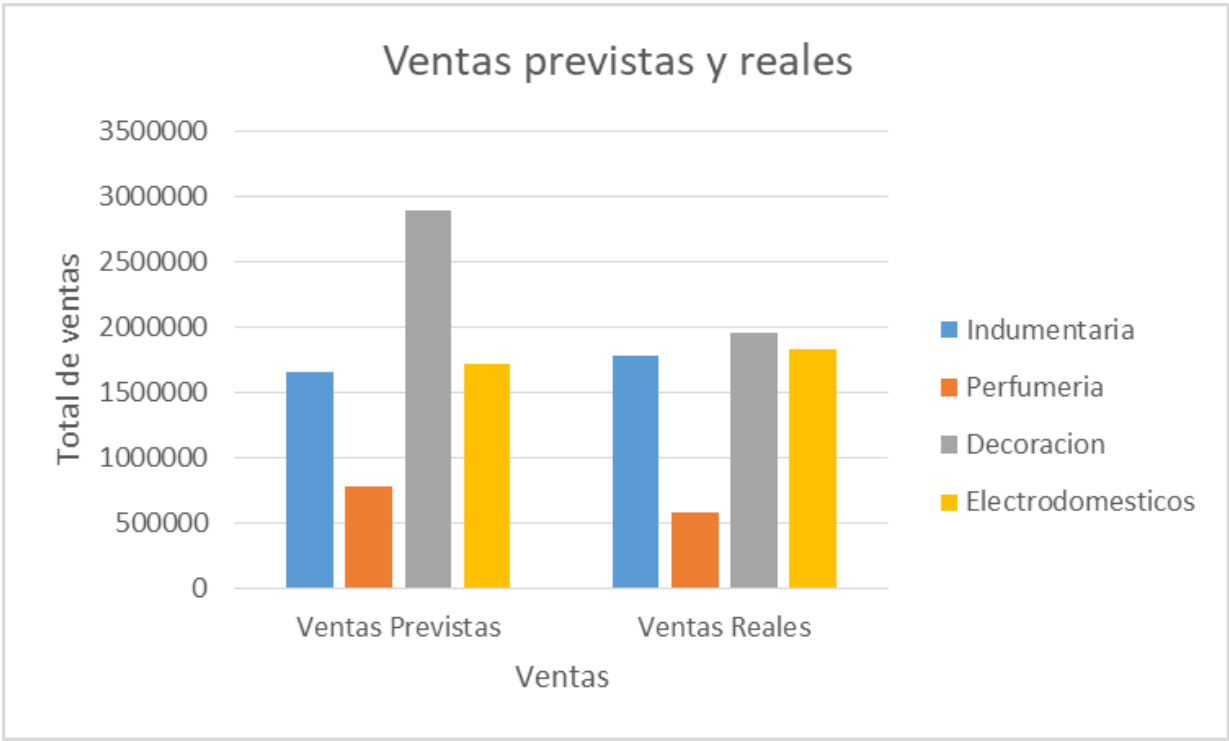
Vendedor	Sector	Numero de clientes	Total de ventas ($)
Constancio	Decoración	620	715000
Gomez	Decoracion	550	660000
Rodriguez	Decoracion	480	582000



4) Ordenar los datos de los **vendedores** alfabéticamente.

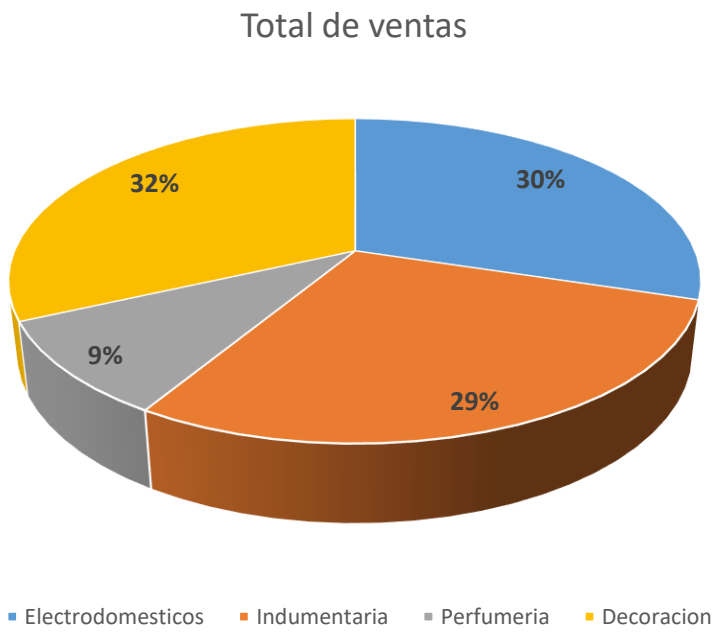
Sector	Ventas Previstas	Ventas Reales
Indumentaria	1650000	1775000
Perfumeria	780000	582500
Decoracion	2890000	1957000
Electrodomesticos	1710000	1830000





5) Ordenar los datos por el total de ventas por área

Sector	Total de ventas
Electrodomésticos	1830000
Indumentaria	1775000
Perfumería	582500
Decoracion	1957000



Ejemplo 1:

Para una aplicación practica de lo antes expuesto, consideremos los siguientes datos correspondientes al diámetro en mm de una cierta pieza derivada de un proceso de producción:

7,33	7,32	7,34	7,4	7,28	7,29	7,35	7,33	7,34	7,34
7,31	7,35	7,32	7,33	7,33	7,36	7,32	7,31	7,35	7,36
7,26	7,39	7,29	7,32	7,34	7,3	7,34	7,32	7,39	7,3
7,33	7,33	7,35	7,34	7,33	7,36	7,33	7,35	7,31	7,33
7,37	7,38	7,38	7,33	7,35	7,3	7,31	7,33	7,35	7,33
7,27	7,33	7,32	7,31	7,34	7,32	7,34	7,32	7,31	7,36
7,3	7,37	7,33	7,32	7,31	7,33	7,32	7,3	7,29	7,38
7,33	7,35	7,32	7,33	7,32	7,34	7,32	7,34	7,32	7,33

1. Construir una distribución de frecuencias de variable continua adecuada a los datos.
2. Representar gráficamente el histograma de frecuencias absolutas simples.
3. Representar gráficamente la poligonal de frecuencias acumuladas
4. Representar gráficamente la poligonal de frecuencias simples

Ejemplo 2:
Dado los siguientes datos correspondientes a los pesos (en kg), del contenido de bolsas de papas:

32,1	31	31,7	31,4	33	33,1	32,6	33	32,4	32,3
34	33	31,8	32,2	34,2	30	31	31,6	31,4	31,3
32,8	32,3	32,7	32,4	29,6	31,4	32,6	34	41,4	32,7
32	32	33,2	34	33	30	31,4	33,1	33,4	31,4
30,1	30,2	33,7	32,7	32,6	31,8	32	31	32,3	33

a) Construir una distribución de frecuencias de 5 intervalos de clase considerando los extremos superiores de los intervalos de clase.

b) Construir el histograma de frecuencias absolutas simples y la correspondiente poligonal de frecuencias acumuladas.

c) Construir la poligonal de frecuencias absolutas simples ¡que estructura tiene la misma?

- **Diagrama de tallo y hoja**

El mismo se al separar los dígitos de cada numero de los datos en dos grupos, un tallo y una hoja. Los dígitos del externo izquierdo son el tallo y están formados por los dígitos de mas alto valor. Los dígitos del extremo derecho son las hojas y contienen los valores mas bajos. Si un conjunto de datos tiene solo dos dígitos, el tallo es el valor de la izquierda y la hoja el de la derecha.

Por ejemplo, si 56 es uno de los números, el tallo es 5 y la hoja 6. Para números con mas dígitos, queda librado al juicio del investigador la determinación del tallo y hoja.

Ejemplo: los siguientes valores se corresponden con las calificaciones obtenidas por 35 aspirantes a un puesto de trabajo:

86	77	91	60	55
76	92	47	88	67
23	59	72	75	83
77	68	82	97	89
81	75	74	39	67
79	83	70	78	91
68	49	56	94	81

Tallo	Hoja									
2	3									
3	9									
4	7	9								
5	5	6	9							
6	0	7	7	8	8					
7	0	2	4	5	5	6	7	7	8	9
8	1	1	2	3	3	6	8	9		
9	1	1	2	4	7					

Como se observa, podemos decir que un Diagrama de Tallos y Hojas presenta las siguientes ventajas:

a) Puede verse si los valores están en el extremo superior o inferior dentro del rango de valores

b) Los valores de los datos originales sin procesar se mantienen, permitiendo dar una estructura del histograma de frecuencias. Recordar que en una distribución de frecuencias de variable continua se utiliza como representativo de los valores de las variables de cada intervalo a las marcas de clase.

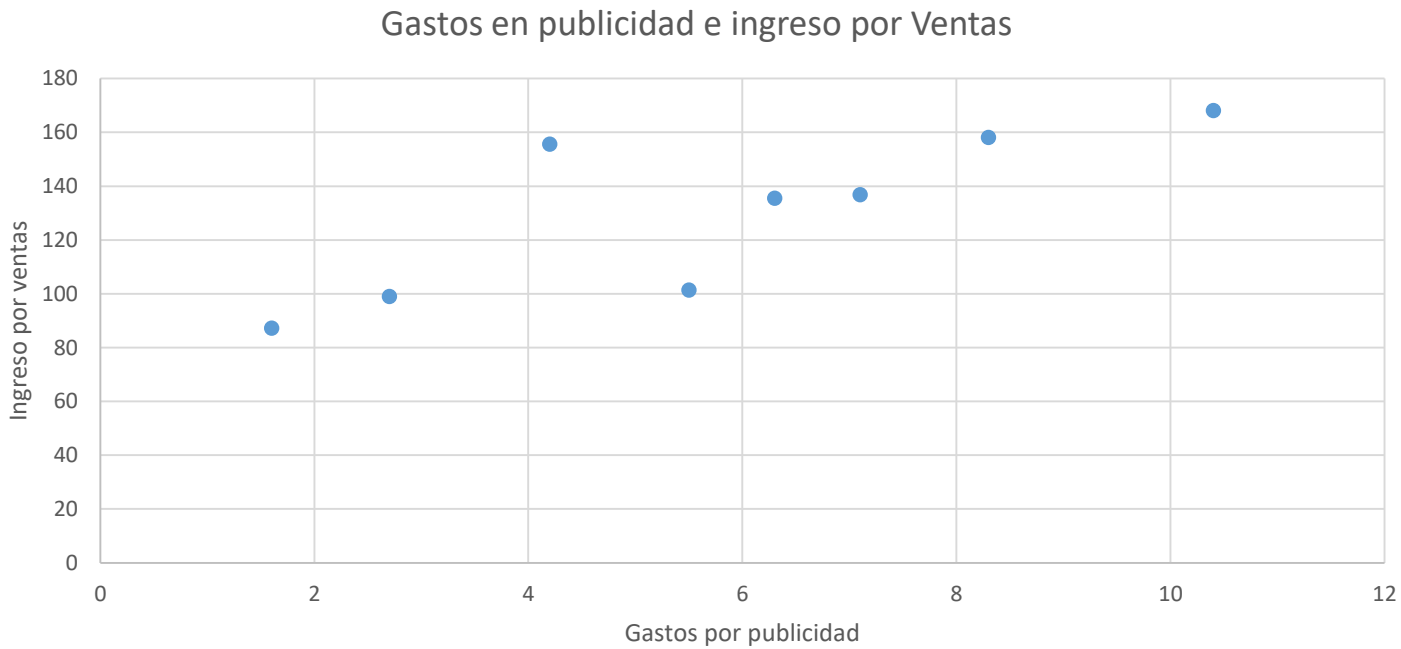
- **Diagrama de dispersión**

En investigaciones de diversa índole, muchas veces es necesario estudiar la relación entre dos variables numéricas. Un mecanismo grafico para ello es el **grafico o diagrama de dispersión o nube de puntos**, dado por un grafico de dos dimensiones conformado por los puntos asociados a dos variables numéricas.

Ejemplo: dados los siguientes datos correspondientes al dinero invertido por varias compañías de una cierta industria en un año reciente en publicidad y los ingresos totales por ventas, ambos en millones de dólares:

Publicidad	Ventas
4,2	155,7
1,6	87,3
6,3	135,6
2,7	99
10,4	168,2
7,1	136,9
5,5	101,4
8,3	158,2

Publicidad	Ventas
4,2	155,7
1,6	87,3
6,3	135,6
2,7	99
10,4	168,2
7,1	136,9
5,5	101,4
8,3	158,2



- **Diagrama de Pareto**

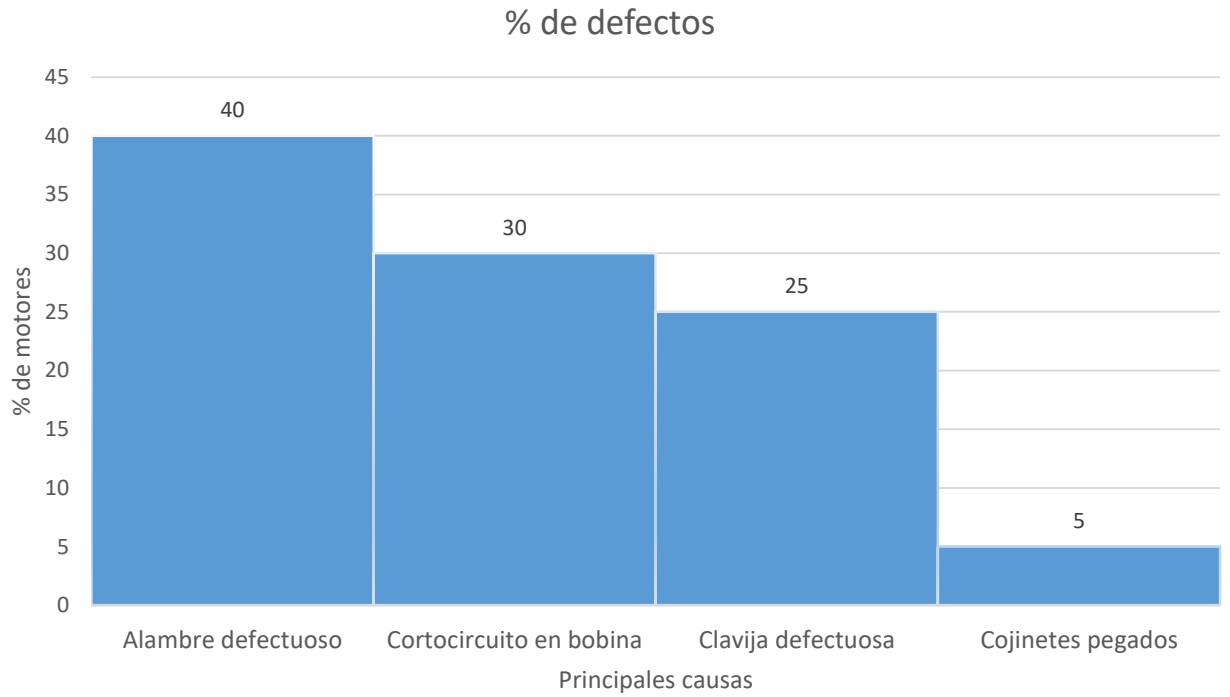
En la actualidad, un concepto importante en la industria es la administración de calidad total la que tiene como finalidad la constante búsqueda de causas de problemas en productos y procesos. Una técnica para mostrar esas causas, es el **análisis de Pareto o grafico de Pareto**, que consiste en un registro cuantitativo del numero y tipo de defectos que se presentan en un producto o servicio. Este grafico de barras verticales que exhibe los tipos de defectos mas comunes, clasificados en el orden de importancia en que se presentan, de izquierda a derecha.

Un grafico de Pareto hace posible separar los defectos mas importantes de los triviales, permitiendo establecer prioridades para quienes deban tomar decisiones en administración de calidad, ya que a veces la mala calidad puede resolverse atacando algunas pocas causas comunes que aparecen en casi todos los problemas.

Ejemplo

El numero de motores eléctricos que son rechazados por los inspectores de calidad de una compañía se ha visto incrementado. Los directivos de la empresa examinan los registros de cientos de motores en los que se encontró por lo menos un defecto, obteniéndose los siguientes datos:

Causa	% de defectos
Alambre defectuoso	40
Cortocircuito en bobina	30
Clavija defectuosa	25
Cojinetes pegados	5



Se observa en el grafico que las principales causas con motores defectuosos son: alambre defectuoso y cortocircuito en bobina, **alcanzando los mismos a un 70 % de todos los problemas.**