UNIVERSIDAD
COMPLUTENSE
MADRID

Convocatoria 2015, Proyecto núm. 164

«Videotutoriales de estadística aplicada a las Ciencias Sociales: un recurso formativo emergente en las actividades de enseñanza-aprendizaje»

Departamento: SOCIOLOGIA IV

Metodología de la Investigación Social y Teoría de la Comunicación

Facultad de Ciencias Políticas y Sociología (UCM)

Estadística y Gráficos

Statistic and graphics

Diagrama de barras, Histograma, Polígono de frecuencias y Gráfico de dispersión

Bar chart, Histogram, Frequencies polygon & Scatter plot

Carlos DE LA PUENTE VIEDMA
Sociología IV – UCM

Índice de la presentación

Introducción: variables, datos y gráficos

Introducción. Variables, datos y gráficos

Estadística:
Trata de la recolección, el análisis, y la presentación de datos organizados en variables.

En la actualidad:
con el **Big data**, las variables pueden acarrear miles de millones de datos. Debido a Internet, la secuenciación de los genomas y los clientes de las grandes empresas.

Gráfico:
Representación de datos numéricos por medio de una o varias líneas que hacen visible la **relación** que esos datos guardan entre sí (RAE).

Representaciones gráficas, que se van a utilizar, según las características de las variables*

	Características de las variables				
	Categóricas		Numéricas		
	Variables		Variables		
Tipo de gráfico	**Una**	**Dos**	**Una**	**2 o más**	**Variación**
Diagrama de barras	**X**	**X**			Tarta o Sectores
Histograma			**X**		Pirámide de Población
Polígono de frecuencias			**X**		Variables estandarizadas
Gráfico de dispersión				**X**	Serie temporal

Fuente: Elaboración propia.
*: La exposición de los gráficos responde a criterios estadísticos, no cronológicos.

Consideramos tres niveles o formatos para presentar los resultados estadísticos: formato de texto (oral y escrito), formato numérico y formato gráfico.

¿Por qué utilizar gráficos?

¿Por qué utilizar gráficos?

Las formas de presentación

La realidad que nos envuelve son **imágenes** y por lo tanto es **gráfica**.

Los seres vivos, y por lo que sabemos, los que tienen **sentido de la vista**, **sistema de la visión** y **sistema nervioso**, se **representan** esta realidad, en forma de **imágenes** y tiene que haber sido así desde el **origen**.

Fuente: (Pedro1267, 2015)

Además, esta representación la **interpretan** y la **entienden** sin necesidad de haber ido a una **escuela de interpretación de imágenes**. Si acaso, la **experiencia**.

Se asume que las **neuronas**, a partir de las **señales** que entran por los ojos, reproducen esas imágenes y **generan**, **almacenan** y **procesan** la **información** asociada a ellas.

La naturaleza nos ha preparado para representarnos la realidad en imágenes. Las neuronas están diseñadas para la representación gráfica.

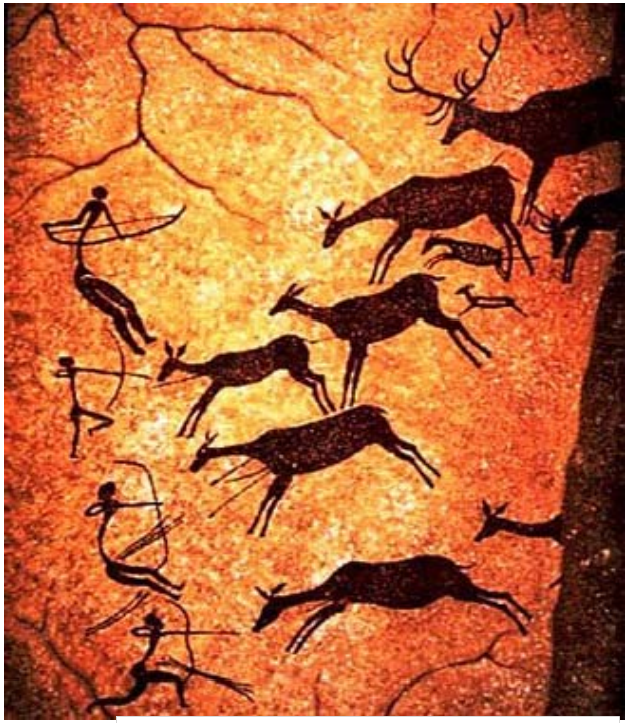
¿Por qué utilizar gráficos?

Las formas de presentación

El ser humano empezó a expresarse **oralmente** hace unos **2 millones de años**. Con sonidos guturales que eran algo más que gruñidos. Asociado al gen **FOXP2** a nivel molecular y a nivel estructural funcional, asociado al aparato fonador y las estructuras corticales.

La **representación gráfica** fue posterior. Ejemplo las pinturas de la Cueva de Altamira.

En el gráfico se puede ver qué animales hay que cazar, cómo hay que cazarlos, que hay más animales que humanos, que se cazan con arco, dónde hay que clavar la flecha, cuántas flechas necesitan lanzar a la presa y la estrategia para cazarlos es de frente. Se puede considerar un gráfico informativo, casi estadístico.



Fuente: (Muñoz & Molero, 2012)

¿Por qué utilizar gráficos?

Las formas de presentación

Posteriormente, mejora la **expresión oral** y **empieza** la **expresión escrita**, pero con caracteres **cuneiformes**. Por lo tanto **gráficos**.

Trozo de tablilla de barro con caracteres cuneiformes.

Fuente: (Muñoz & Molero, 2012)

La **expresión numérica** empieza también con cifras **cuneiformes**.

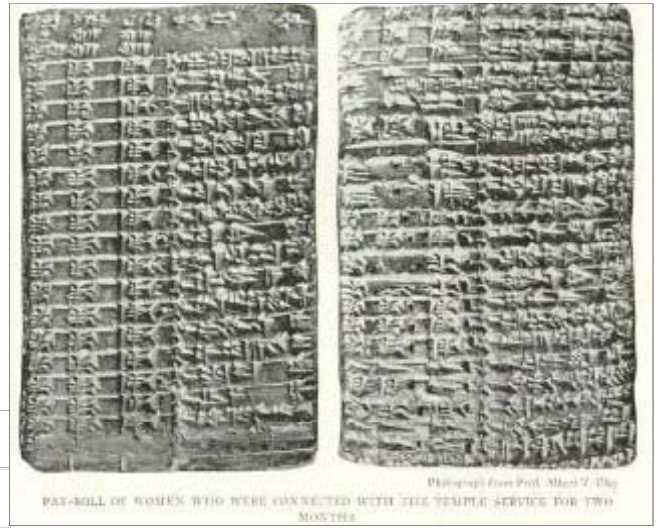


Tabla de multiplicar con cifras cuneiformes.

Fuente: (Herrera Arellano)

¿Por qué utilizar gráficos?

Las formas de presentación

Para los **cálculos numéricos**, los **griegos** utilizaron letras mayúsculas y minúsculas. En la figura uno de los sistemas numéricos que usaron y los valores asociados.

Los **romanos** utilizaron la numeración romana y los **numerales etruscos** pudieron ser sus antecesores.

Los **hindúes** generan el sistema de numeración que después desarrollaron los **árabes** y que utilizamos hoy como **numeración arábiga**.

Letra	Valor	Letra	Valor	Letra	Valor
α΄	1	ι΄	10	ρ΄	100
β΄	2	κ΄	20	σ΄	200
γ΄	3	λ΄	30	τ΄	300
δ΄	4	μ΄	40	υ΄	400
ε΄	5	ν΄	50	φ΄	500
ϝ΄ / ϛ΄ / στ΄	6	ξ΄	60	χ΄	600
ζ΄	7	ο΄	70	ψ΄	700
η΄	8	π΄	80	ω΄	800
θ΄	9	ϙ΄ / ϟ΄	90	ϡ΄	900

Fuente: (Wikipedia, 2015)

El sistema numérico, como lo conocemos hoy, es la forma de comunicación y presentación que ha aparecido en último lugar.

Entonces, la Naturaleza nos ha preparado para representarnos una realidad que son imágenes. Pero los humanos nos tenemos que preparar para la representación y expresión de texto oral y escrito, y numérica.

¿Por qué utilizar gráficos?

Resumiendo

Por todo esto, los resultados estadísticos de **texto** y **numéricos**, **deben** acompañarse con **gráficos** porque es la **forma natural** desde donde se reciben las señales (**fotones**) que se procesan en las **neuronas** como imágenes y se les asocia la información: **significados** y **contenidos**. El **texto** y los **números** aparecen después y tienen que ser **aprendidos**.

No obstante, los **gráficos** se asocian a **imágenes**, pero también tienen **significados**. El **texto** y los **números** se asocian a **significados**, pero también son **imágenes**. Todo son imágenes que se reproducen en las **neuronas** a través de los **fotones** que entran por los ojos.

Por lo tanto, los **números**, las **palabras**, las **sílabas**, las **letras** y los propios **gráficos**, no son otra cosa que **imágenes** que proyectan **paquetes de fotones contorneados** y asumiendo que los fotones son lo que son y no acarrean información, el **contenido** y **significado** lo debe proporcionar lo que llamamos la **consciencia**.

Quien muestra o expone los gráficos debe controlar las funciones y fórmulas estadísticas para expresarlos de forma oral, escrita y numérica.

Quien observa los gráficos, al menos, debe conocer la forma de expresión oral y numérica, para tener una mejor comprensión de lo que se le expone y observa.

Datos que se van a utilizar y representar

Datos que se van a utilizar y representar

Los **datos** que se van a representar gráficamente están aglutinados en **variables** y pueden tener tres **formatos**.

Conocidos como tipo de datos: **T-I**, **T-II** y **T-III**.

Datos **T-I**, matriz de datos o micro datos.

Datos **T-II** o tabla de frecuencias.

Datos **T-III** o tabla de frecuencias agrupada por intervalos.

Los datos **T-I**, si no se opera con un programa estadístico, hay que pasarlos a **T-II** o **T-III**.

Parte de Matriz de datos o micro datos (T-I)

idno	tvtot	tvpol	pplrtrst	pplfair	pplhlp	polintr	psppsgv	actrolg	psppipl
1	4	1	7	7	8	2	0	4	2
2	7	3	5	5	3	1	5	5	1
3	6	2	6	8	7	2	3	0	0
4	3	1	5	3	2	2	3	7	4
5	2	2	3	7	8	4	0	0	0
6	2	2	0	10	5	1	1	5	0
7	7	5	5	6	7	2	2	6	2
13	3	1	5	7	4	3	1	6	3
14	4	1	9	6	3	2	4	2	5
21	5	2	5	4	7	4	3	0	3
22	3	1	3	5	5	3	3	1	2
23	0	66	9	8	6	3	6	2	4
24	7	1	5	5	5	4	5	1	7
25	5	1	5	5	3	3	4	1	4
26	4	3	4	4	5	3	7	6	6
33	4	3	5	4	4	2	2	6	3
34	3	1	3	2	0	2	0	5	3
35	4	1	5	7	7	3	6	0	3
36	3	2	6	5	6	2	2	4	2
37	6	3	6	9	9	2	5	6	8
38	4	1	6	5	6	3	4	0	3
39	1	1	9	2	9	2	8	8	7
40	6	1	5	2	6	4	0	0	0
45	7	1	7	9	10	2	5	2	6

Fuente: ESS round 7.
Tabla: Elaboración propia.

Tabla de datos T-II

Entrevistados según el sexo

Sexo		F. absoluta (n)	F. relativa (%)
	Varón	1.205	48,5
	Mujer	1.281	51,5
	Total	**2.486**	**100,0**

Fuente: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.

Tabla de datos T-II

Entrevistados según la edad
(Ver diapositiva 39)

Edad		F. absoluta (n)	F. relativa (%)
	18	22	0,9
	19	34	1,4
	20	25	1,0
	[...]	[...]	[...]
	93	2	0,1
	94	1	0,0
	Total	**2.486**	**100,0**

Fuente: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.

Tabla de datos T-III

Entrevistados según la edad

Edad		F. absoluta (n)	F. relativa (%)
	15 - 25	243	9,8
	25 - 35	370	14,9
	35 - 45	493	19,8
	45 - 55	467	18,8
	55 - 65	367	14,8
	65 – 75	309	12,4
	75 - 85	191	7,7
	85 - 95	46	1,9
	Total	**2.486**	**100,0**

Fuente: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.

Sistema de representación de los gráficos

Sistema de representación de los gráficos

La representación gráfica se va a realizar en un sistema de **coordenadas cartesianas** de dos dimensiones.

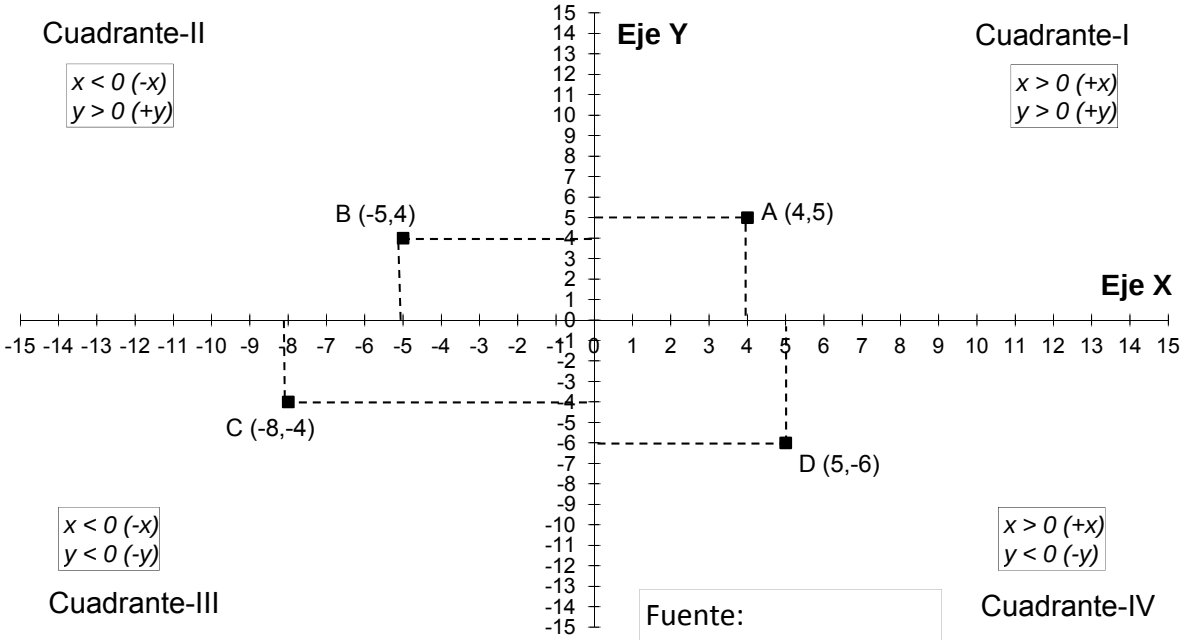
Son dos ejes que se cruzan perpendicularmente y definen un plano en el que forman cuatro cuadrantes: **Cuadrante-I**, **Cuadrante-II**, **Cuadrante-III** y **Cuadrante-IV**.

En el **eje horizontal** o de abscisas se representa una variable, normalmente llamada *X*. En el **eje vertical** o de ordenadas se representa otra variable, normalmente llamada *Y*.

El punto en donde se cortan los ejes se denomina el "**punto de origen de las coordenadas**" y tienen el valor 0 para la *X* y para la *Y*. Por lo tanto se representa como punto (*x, y*), de valores (0, 0). Primero se designa la *X* y después la *Y*.

A partir de este planteamiento, se **escala cada eje** de acuerdo a la variable que se quiere representar en él.

Sistema de dos ejes (Cuatro cuadrantes)



Fuente:
De la Puente, 2015

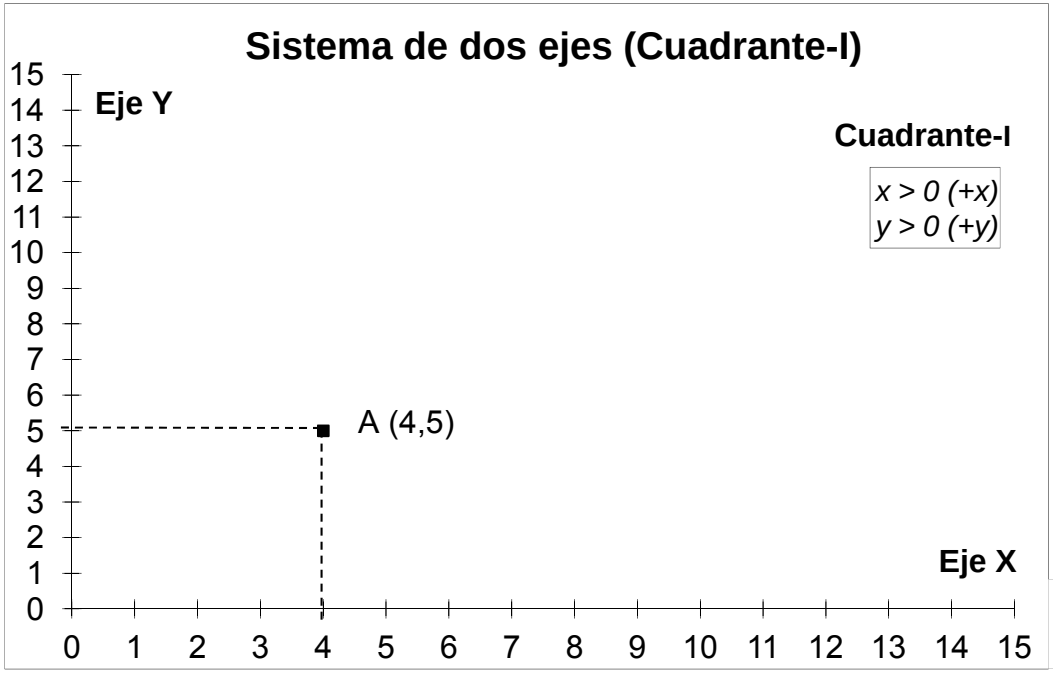
Cada **unidad** de la variable se representa por un **segmento de forma arbitraria**, de tal manera que el eje represente desde el valor mínimo al máximo de la variable y que se ajuste a las dimensiones de la caja del gráfico y dentro de las dimensiones de la hoja que se utiliza.

Por lo tanto, en el eje *X*, desde el punto de origen a la derecha toma valores **positivos** y hacia la izquierda, valores **negativos**. De la misma manera, el eje *Y* toma valores **positivos** hacia arriba desde el punto cero de origen, y valores **negativos** hacia abajo.

Una vez así dispuestos los ejes, se puede representar cualquier punto en el plano en base a los valores en dos variables, que se trasladan a la escala en sus respectivos ejes.

Sistema de representación de los gráficos

Si las variables tienen los valores en el rango de los **valores positivos**, entonces se trabaja en el **Cuadrante-I**. Esto es, con valores positivos en el eje horizontal, de abscisa o *X*, y en el eje vertical, de ordenadas o *Y*.



Fuente:
De la Puente, 2015

Sistema de representación de los gráficos

Presentación y lectura de los gráficos

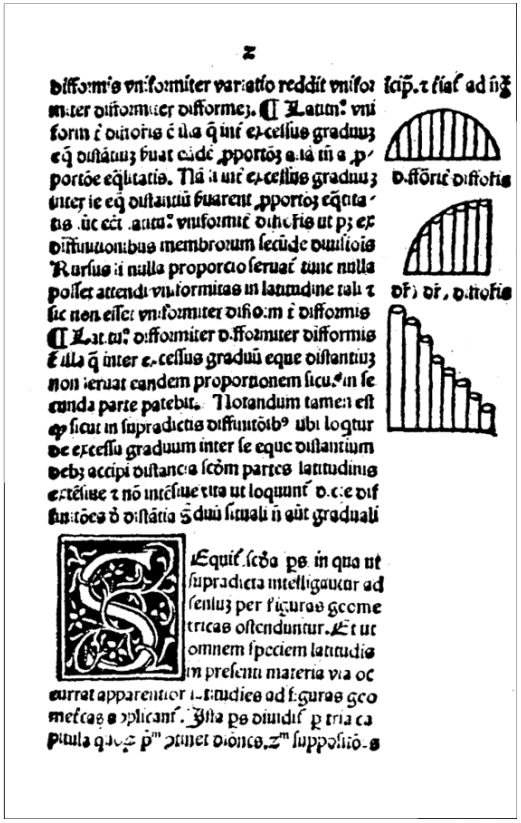
La presentación de un gráfico tiene tres niveles:

1. El gráfico propiamente dicho, con los títulos
2. El nivel numérico de los valores que se representan en el gráfico.
3. El nivel textual, hablado y/o escrito, para expresar el significado del gráfico

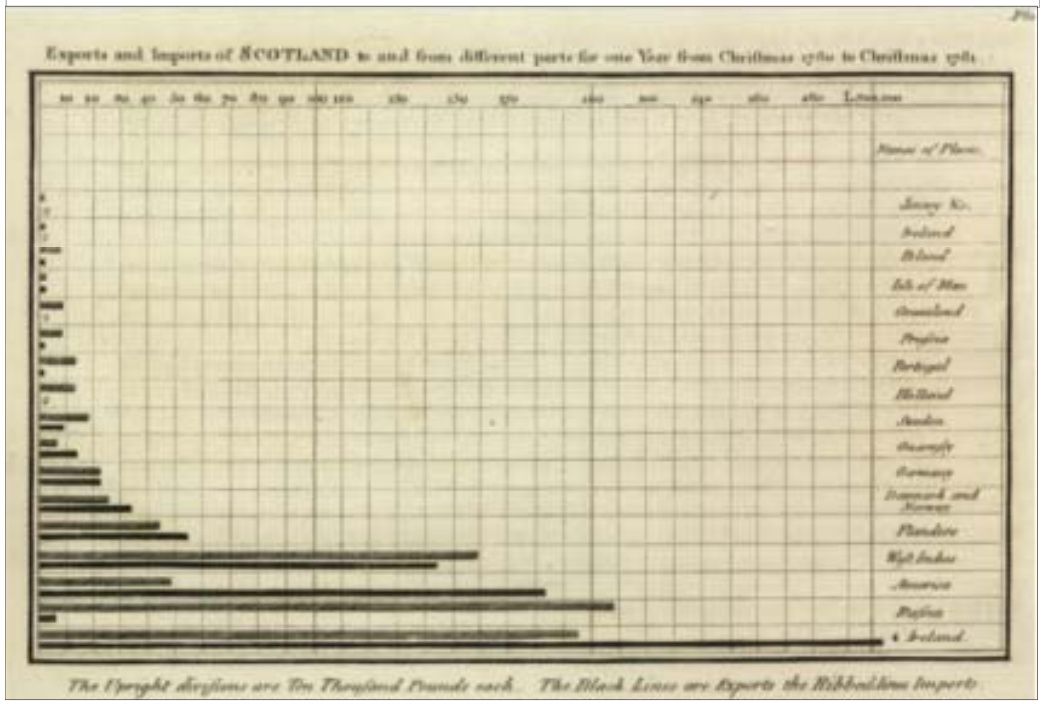
Antecedentes de los gráficos estadísticos

Antecedentes de los gráficos

Hay constancia de un **diagrama de barras** o **posible histograma** de **Nicolas Oresme** en el siglo XIV (Wikipedia, 2015b).

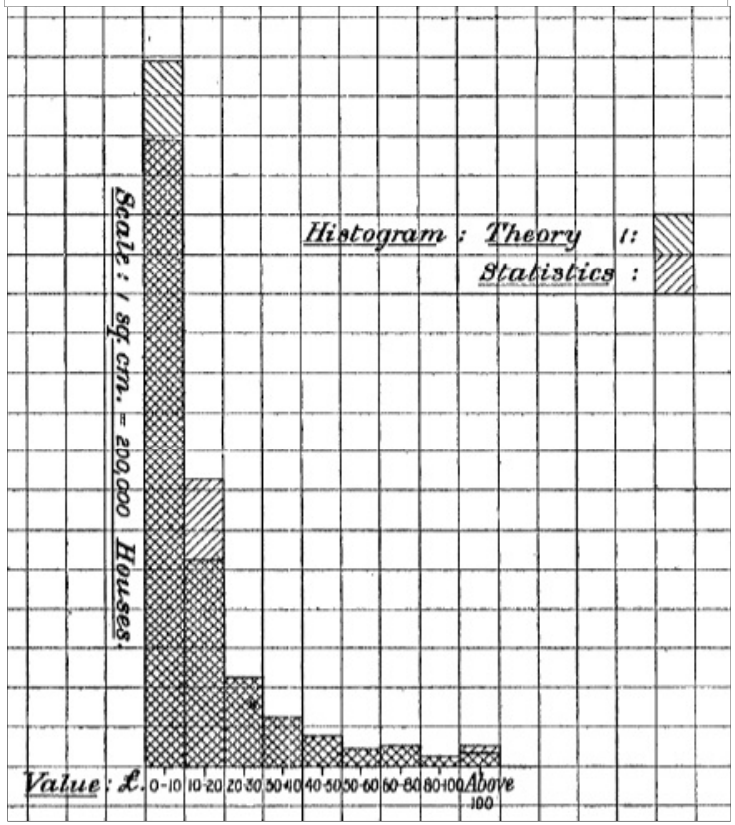


Un trabajo pionero en la representación gráfica es el de **William Playfair** (1759-1823) que en 1786 utilizó un **diagrama de barras** (Wikipedia, 2015c). También utilizó **otros gráficos** como **series temporales** y **gráficos de tarta** (Spence, 2006).

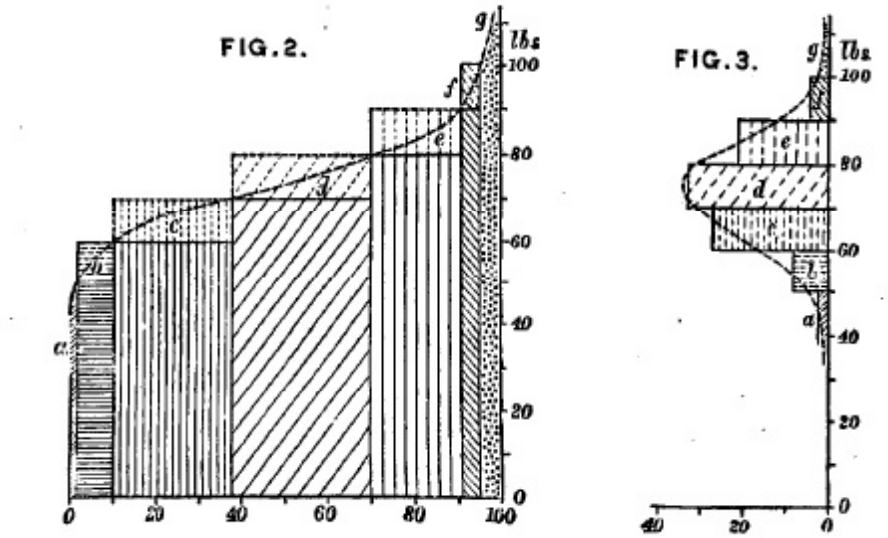


Antecedentes de los gráficos

Karl Pearson nombra el **histograma** como tal en una publicación de 1895 (Pearson, 1895).

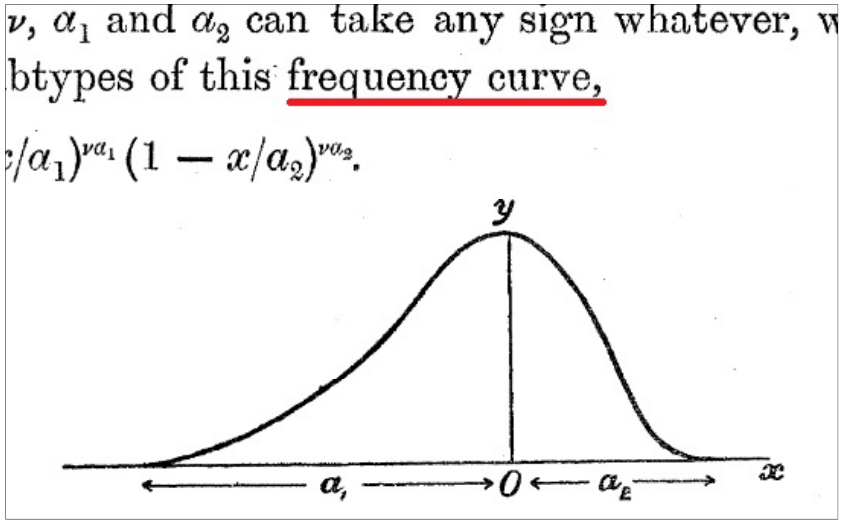


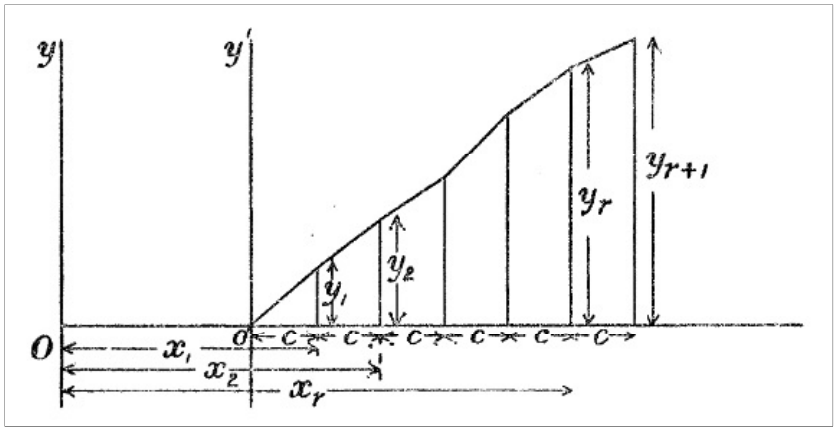
En **Francis Galton** también aparecen el **histograma**, el **polígono de frecuencias** y el **polígono de frecuencias acumuladas** (Galton, 1889).



Antecedentes de los gráficos

En la misma publicación, **Pearson**, también hace referencia al **polígono de frecuencias** y al **polígono de frecuencias acumuladas** (Pearson, 1895).





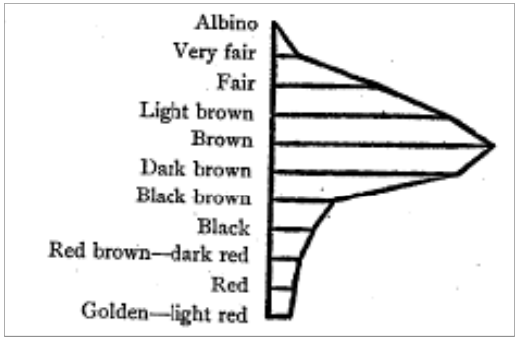
Antecedentes de los gráficos

John Graunt elabora **tablas estadísticas** separadas para **varones** y **mujeres** (Graunt, 1662/1996) y se puede considerar la base de la **pirámide de población**.

The Table of Males and Females

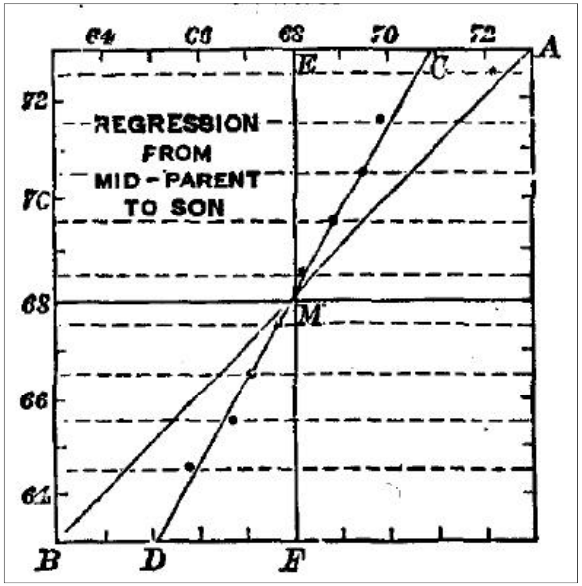
Years	Commu-nicants	Wed-dings	Christned M.	Christned F.	Christned Both	Buried M.	Buried F.	Buried Both
1589		20	31	27	58	28	16	44
90		16	40	29	69	36	21	57
91		12	37	28	65	35	30	65
92		14	40	25	65	28	19	47
93		20	32	20	52	33	32	65
94		24	34	37	71	16	22	38
95		16	32	28	60	33	28	61
96		9	36	26	62	42	29	71
97		23	23	25	48	53	64	117
98		21	37	29	66	33	23	66
		175	342	274	616	337	284	631

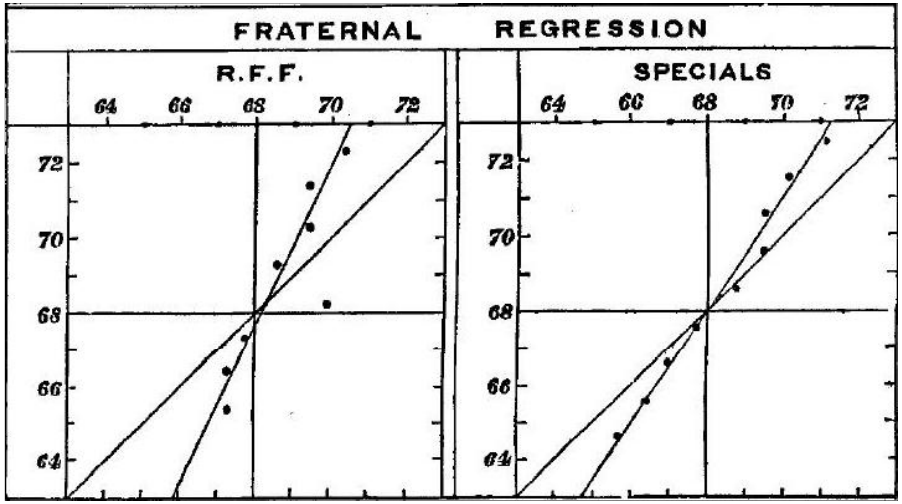
Este gráfico de **Francis Galton** y los anteriores se pueden considerar **indicios** de la **pirámide de población** (Galton, 1889).



Antecedentes de los gráficos

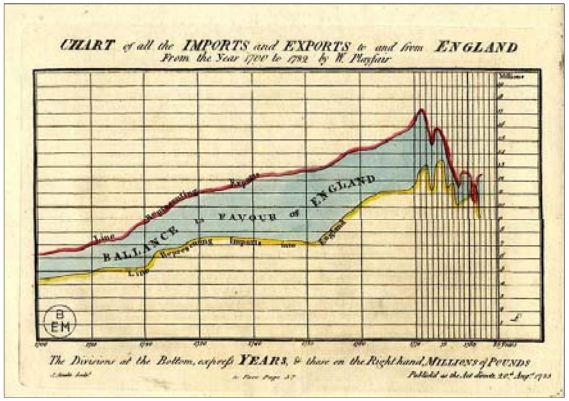
El **gráfico de dispersión** se le atribuye a **Francis Galton** (Wikipedia, 2015d). En su texto “Natural Inheritance” Galton muestra estos gráficos (1889)

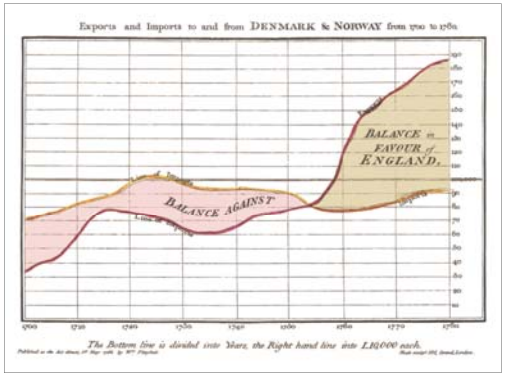




Antecedentes de los gráficos

Aunque hay ejemplos de **series temporales** en el **siglo X**, la forma más clara son las **series temporales** de **William Playfair** en 1786 (Wikipdeia, 2015a). Se muestran algunos ejemplos.





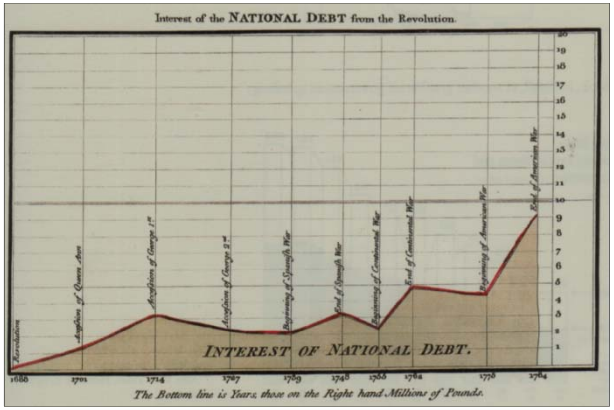


Diagrama de barras

Diagrama de barras

El **diagrama de barras** es la representación gráfica de variables **categóricas** en formato de datos tipo **T-II** o tabla de frecuencias, en un sistema de coordenadas cartesianas de dos dimensiones. Como los valores son positivos, la representación se hace en el cuadrante-I.

Se representa la tabla de la variable "**Sexo de la persona entrevistada**" del Barómetro (Nº 3104) del CIS, realizado en el mes de julio de 2015.

Diagrama de barras

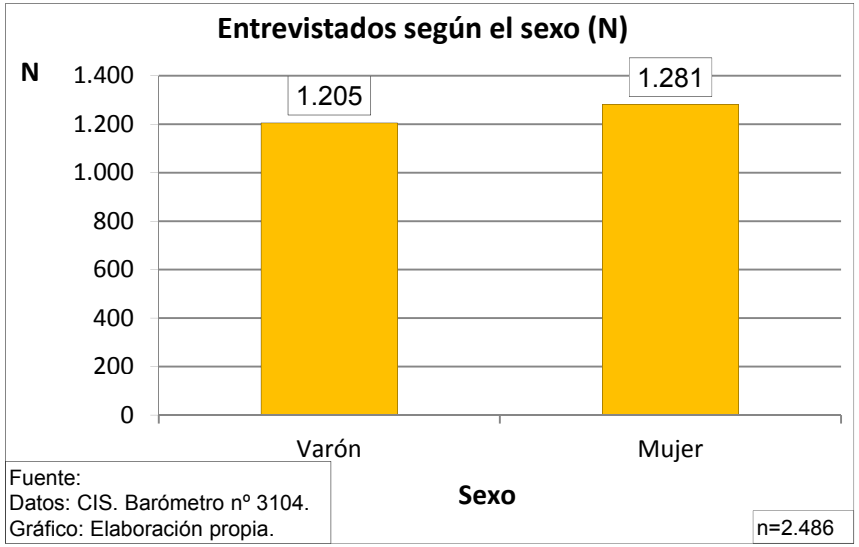
En el **eje horizontal** o *X* se representa las categorías o sucesos elementales de la variable "**Sexo**". Como no existe relación entre las categorías, las marcas se sitúan aleatoriamente en el eje y las columnas no tienen contacto entre ellas porque no hay continuidad. En el **eje vertical** o *Y* se representa el **número de casos** (Frecuencias absolutas) o **porcentajes** (Frecuencias relativas).

Nivel numérico

Entrevistados según el sexo			
Sexo		**F. absoluta (N)**	**F. relativa (%)**
	Varón	1.205	48,5
	Mujer	1.281	51,5
	Total	**2.486**	**100,0**

Fuente: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.

Nivel gráfico



El diagrama en formato de columnas muestra los valores absolutos de varones y mujeres y la relación visual que hay entre las dos categorías.

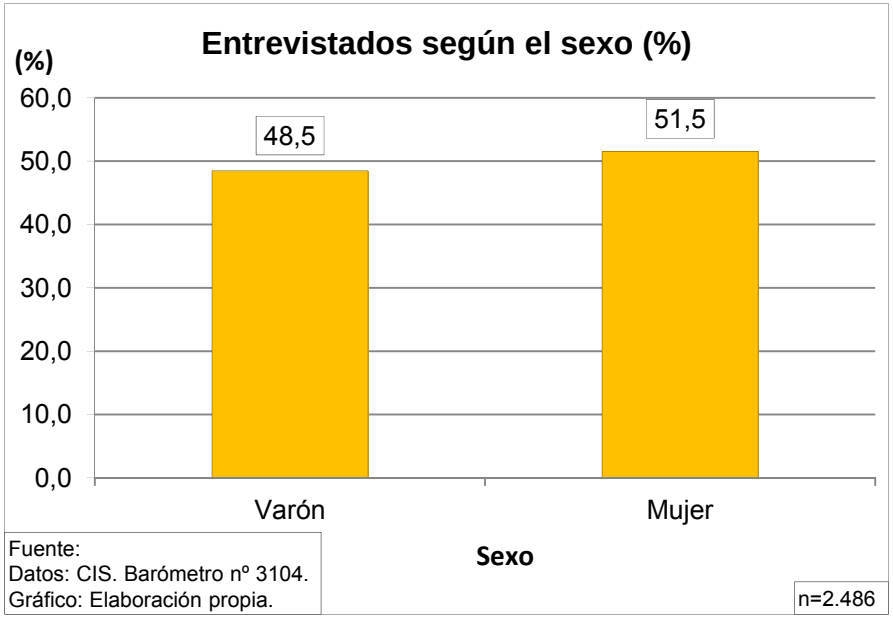
[**Nivel texto**] Una lectura sería: "**El número de mujeres (1.281) es mayor que el de varones (1.205)**". El nombre del gráfico es el de la variable que se representa. Pudiendo recoger la característica de los valores representados, que en este caso son valores absolutos.

Diagrama de barras

Al expresarlo en **porcentajes**, la relación gráfica es la misma que con los valores absolutos, pero muestra la diferencia de proporción, y en el caso de ser una muestra representativa se puede tener idea de magnitud y de la diferencia en la población.

Entrevistados según el sexo			
Sexo		**F. absoluta (N)**	**F. relativa (%)**
	Varón	1.205	48,5
	Mujer	1.281	51,5
	Total	**2.486**	**100,0**

Fuente: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.



Una lectura sería: "**Las mujeres son el 51,5 % y los varones el 48,5%, por lo tanto, la diferencia es de 3 puntos porcentuales**". Si no se indica otra cosa, los porcentajes se presentan con un decimal.

Diagrama de barras

El diagrama de barras, en formato columna, se puede presentar en formato de **barras** o como diagrama de **tarta** o **sectores**.

Las lecturas son las mismas.

Diagrama de barras en formato de barras

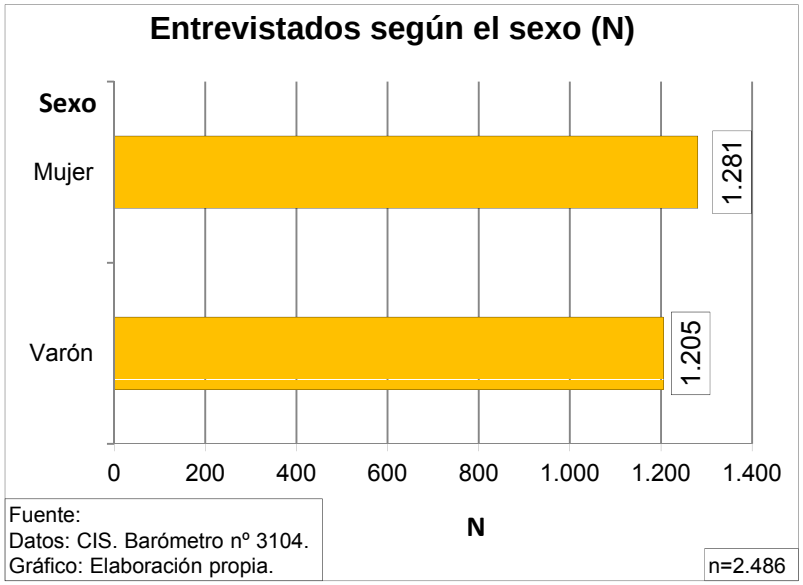


Diagrama de tarta o sectores

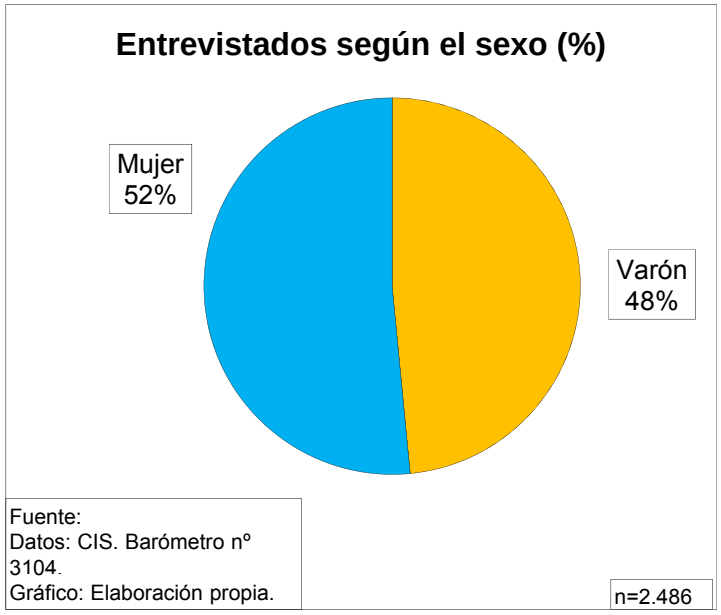


Diagrama de barras

Con dos variables

Con el diagrama de barras o columnas se puede representar **más de una variable**. En el caso de dos variables, una de ellas se representaría en función de o dentro de las categorías de la otra. En este caso se van a utilizar la variable **sexo** y **el nivel de estudios**. La representación puede ser, el **sexo según el nivel de estudios** o también **el nivel de estudios según el sexo**. El origen de los datos es una tabla de doble entrada.

Fuente:
Datos: CIS. Barómetro nº 3104.
Tabla: Elaboración propia.

Nivel de estudios por Sexo de la persona entrevistada

			Sexo de la persona entrevistada						Total
			Varón			Mujer			
Nivel de estudios	S/estudios	N	64			90			154
		% TF/TC/TT	41,6%	5,3%	2,6%	58,4%	7,0%	3,6%	
	Primaria	N	196			251			447
		% TF/TC/TT	43,8%	16,3%	7,9%	56,2%	19,6%	10,1%	
	Secundaria	N	308			301			609
		% TF/TC/TT	50,6%	25,6%	12,4%	49,4%	23,5%	12,1%	
	Bachiller	N	156			154			310
		% TF/TC/TT	50,3%	13,0%	6,3%	49,7%	12,0%	6,2%	
	FP	N	245			204			449
		% TF/TC/TT	54,6%	20,3%	9,9%	45,4%	15,9%	8,2%	
	Superiores	N	235			279			514
		% TF/TC/TT	45,7%	19,5%	9,5%	54,3%	21,8%	11,2%	
Total		N	1204			1279			2483

Diagrama de barras

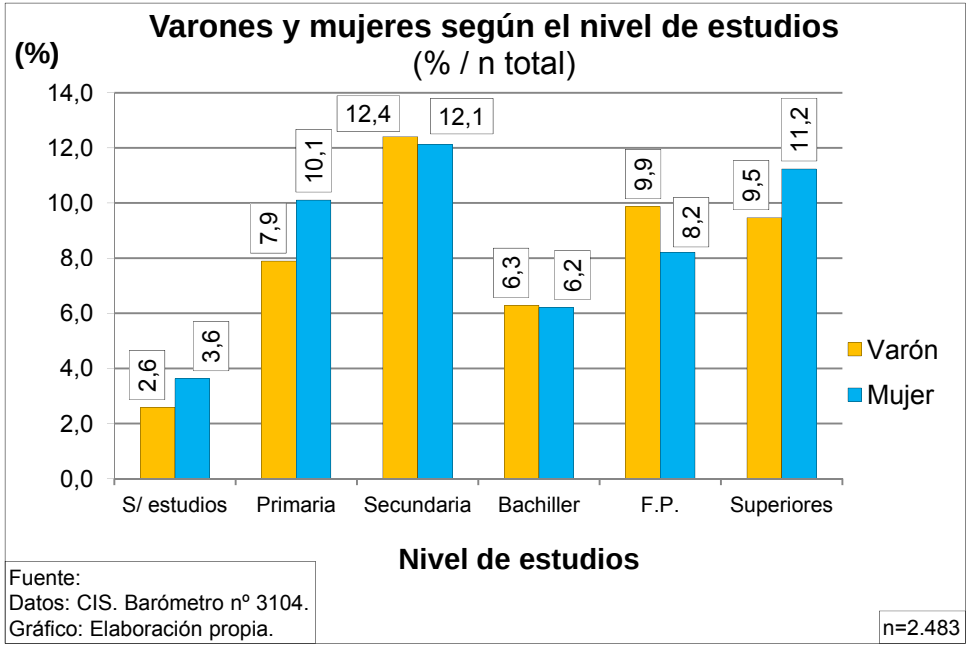
Con dos variables

Diagrama de barras en formato columna **con dos variables**, según la tabla de doble entrada anterior. En la representación del **sexo** según el **nivel de estudios**, las **comparaciones** o lectura sería la relación entre **varones y mujeres** dentro del mismo nivel de estudios, y genera dos **series**, la de "**Varón**" y la de "**Mujer**".

Este tipo de gráficos tiene diversas lecturas.

- La **exhaustiva**, que sería leer de forma monótona los porcentajes de todas las barras. De esta manera no se aporta nada al gráfico.
- Hacer referencia a lo que **más destaca**.
- Hacer referencia a lo que **menos destaca**.
- Comparar las **diferencias mayores**.
- Comparar las **diferencias menores**.

El interés de la lectura estará en función del tema o asunto del que se trate.



Una lectura: "**El 9,9 % son varones que estudian FP, frente al 8,2% que son mujeres. Pero en estudios superiores las mujeres presentan un porcentaje mayor (11,2%) que los varones (9,5%)**".

Otra lectura: "**El porcentaje mayor en el nivel de estudios, tanto en varones como en mujeres, es en 'Secundaria', 12,4 % y 12,1 % respectivamente**".

Diagrama de barras

Con dos variables

Si se cambia la base para el cálculo de los porcentajes, las cantidades presentan variaciones. Presenta una lectura distinta, pero también es correcta.

Por ejemplo, los porcentajes de varones y mujeres sin estudios son mayores que en la tabla anterior, porque la base es menor. Es el total de entrevistados sin estudios (n=154).

Una lectura: "**Respecto del total en cada nivel de estudios, el porcentaje mayor es el de mujeres "Sin estudios" (58,4%). Pero respecto del total de la muestra son el 3,6%**". (Ver el gráfico anterior).

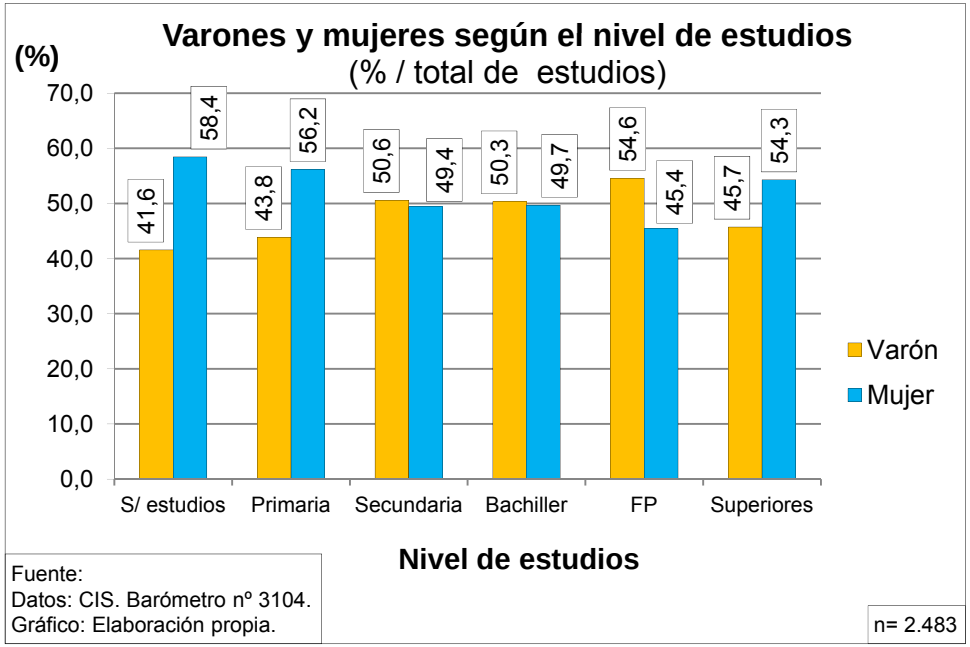


Diagrama de barras

Con dos variables

Si se considera el total de varones y mujeres, el punto de vista cambia y los porcentajes también. Todas las lecturas son correctas pero dan una información distinta. Otro punto de vista.

Una lectura: "**El 20,3% de los varones estudia FP, mientras que las mujeres son el 15,9%. La diferencia es de 4,4 puntos porcentuales**". Al considerar el total de la muestra, la diferencia era de tres puntos porcentuales.

Otra lectura: "**Los porcentajes mayores de varones y mujeres por nivel de estudios se da en 'Secundaria' , que son el 25,6% y 23,5%, respectivamente"**

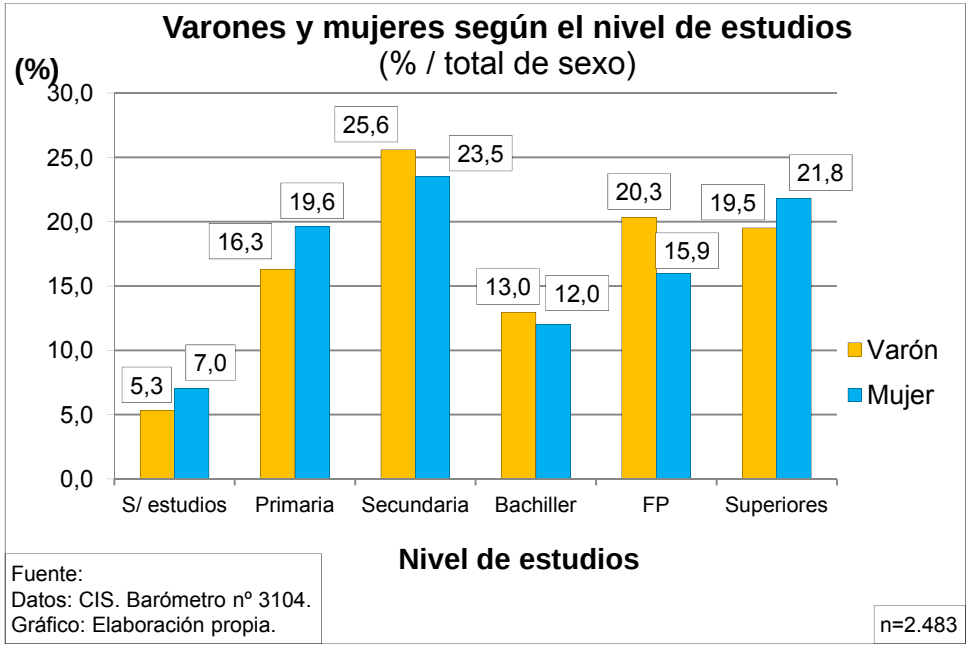
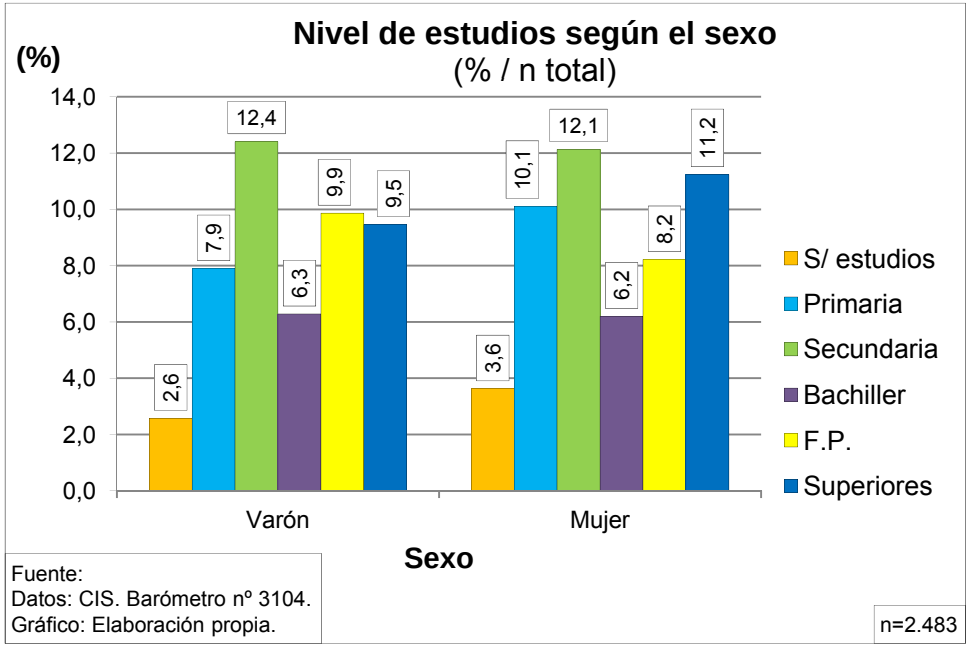


Diagrama de barras

Con dos variables

La representación del **nivel de estudios** dentro de cada categoría de **sexo**, produciría este gráfico. Los porcentajes son los mismos pero agrupados de otra manera.

Este gráfico facilita la **comparación** entre los distintos niveles de estudios, dentro de cada categoría de sexo. Como en el gráfico sobre **n** total, el porcentaje mayor se da en el nivel de estudios de secundaria, 12,4% son varones que tienen estudios de secundaria y el 12,1% son mujeres.



Histograma de frecuencias

Histograma de frecuencias

El **histograma** o **histograma de frecuencias** es la representación gráfica de una **variable numérica** en formato de datos tipo **T-II** o **T-III**, en un sistema de coordenadas cartesianas de dos dimensiones. Si los valores son positivos, se utiliza sólo el cuadrante-I. En el **eje *X*** se representa la **variable** y se define una longitud de segmento unitario por unidad de la variable. En el **eje *Y*** se representan el **número de casos** (frecuencias absolutas) o el **porcentaje de casos** (frecuencia relativa).

El origen de los datos puede ser también una matriz de datos o micro datos (**tipo T-I**). Pero si no se opera con un programa estadístico, hay que transformarlos en **T-II** o **T-III**.

Los **intervalos** pueden ser de **igual** o **distinta amplitud**. El **número de intervalos** puede estar determinado por:

- Fórmula
- Criterios de interés
- Un criterio razonable
- Percentiles cuartiles, u otros

Las **fórmulas** pueden ser, la regla de **Sturges** o la regla de la **raíz cuadrada**.

En un **criterio de interés** los intervalos estarían definidos por el estudio o **marco teórico** de referencia.

El **criterio razonable** es que los intervalos sean entre **5 y 15**, dependiendo también del número de casos “n”.

El criterio de interés en **edad** o de **tamaño de municipios** puede ser utilizar el del **CIS** (Centro de Investigaciones Sociológicas).

En esta ocasión, la variable edad de la persona entrevistada se ha agrupado en intervalos de 10 años de amplitud.

Histograma de frecuencias

La variable que se va a representar es "**Edad de la persona entrevistada**" del Barómetro (Nº 3104) del CIS realizado en el mes de julio de 2015.

Tabla de datos Tipo-II

Entrevistados según la edad		
	F. absoluta	F. relativa
18	22	0,9
19	34	1,4
20	25	1,0
21	28	1,1
22	31	1,2
23	26	1,0
24	38	1,5
25	39	1,6
26	29	1,2
27	32	1,3
28	37	1,5
29	31	1,2
30	49	2,0
31	32	1,3
32	35	1,4
33	36	1,4
34	40	1,6
35	49	2,0
36	41	1,6
37	39	1,6
38	38	1,5
39	52	2,1
40	59	2,4
41	52	2,1
42	53	2,1
43	62	2,5
44	50	2,0

Entrevistados según la edad		
	F. absoluta	F. relativa
45	47	1,9
46	38	1,5
47	51	2,1
48	48	1,9
49	41	1,6
50	42	1,7
51	52	2,1
52	42	1,7
53	58	2,3
54	54	2,2
55	41	1,6
56	40	1,6
57	35	1,4
58	32	1,3
59	35	1,4
60	37	1,5
61	36	1,4
62	36	1,4
63	39	1,6
64	41	1,6
65	36	1,4
66	35	1,4
67	34	1,4
68	34	1,4
69	37	1,5
70	36	1,4

Entrevistados según la edad		
	F. absoluta	F. relativa
71	28	1,1
72	30	1,2
73	27	1,1
74	23	0,9
75	25	1,0
76	14	0,6
77	23	0,9
78	27	1,1
79	26	1,0
80	24	1,0
81	13	0,5
82	22	0,9
83	20	0,8
84	14	0,6
85	8	0,3
86	6	0,2
87	14	0,6
88	9	0,4
89	6	0,2
90	4	0,2
91	2	0,1
92	2	0,1
93	2	0,1
94	1	0,0
Total	2486	100,0

Fuente:
Datos: CIS. Barómetro nº 3104
Tabla: Elaboración propia

Histograma de frecuencias

La variable se agrupa en intervalos de 10 años de amplitud, empezando en 15 años y resultan 8 intervalos, como criterio razonable entre 5 y 15 intervalos y coincide con la fórmula de Sturges. El rango de la variable va de la edad mínima 18 años hasta la máxima de 94. Pero para equilibrar e igualar los intervalos, se empieza desde 15 años y se termina en 95 años.

$$c = 1 + 3{,}322 + log_{10}(n) = 1 + 3{,}322 + log_{10}(2.486) = 7{,}72 = 8\ intervalos$$

La tabla resultante es de tipo **T-III**. Por ejemplo, los 493 individuos que tienen 35 años, en realidad tendrán 35 o más pero menos de 45, por lo que se consideran distribuidos a lo largo del intervalo, y así para todas las categorías.

Fuente:
Datos: CIS. Barómetro nº 3104
Tabla: Elaboración propia

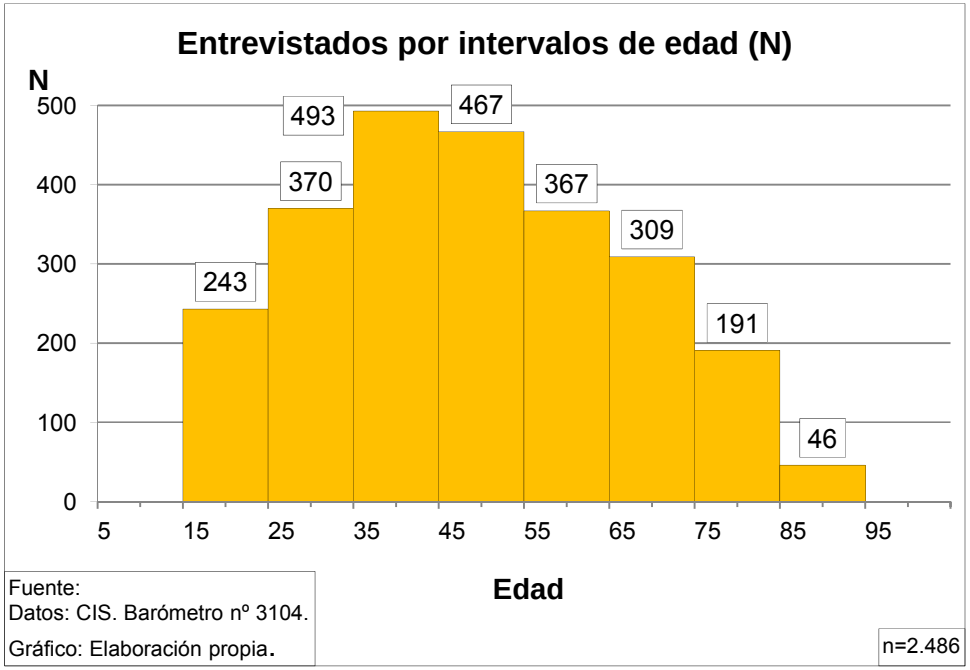
Entrevistados según la edad

Edad		F. abs. (n)	F. rel. (%)
	15-25	243	9,8
	25-35	370	14,9
	35-45	493	19,8
	45-55	467	18,8
	55-65	367	14,8
	65-75	309	12,4
	75-85	191	7,7
	85-95	46	1,9
	Total	**2.486**	**100,0**

Histograma de frecuencias

Así planteado, el histograma de frecuencias absolutas muestra el **número de casos que hay por intervalo de edad** y la lectura sería "**En el intervalo de 15 a 25 años, hay 243 casos**". Esto es, que tienen 15 o más años pero menos de 25 años. "**De 25 a 35 años, hay 370 casos**". Esto es, que tienen 25 años o más pero menos de 35 años, y así sucesivamente para todos los intervalos. Como la variable es numérica, entonces las columnas son contiguas.

Como las columnas tienen igual ancho, si se considera que tiene valor unitario, al multiplicarlo por la altura, que es la frecuencia, la **superficie** de cada columna representa al **número de casos**. Por lo tanto, la **superficie total** del histograma representa al **total de casos**. Si se considera las frecuencias relativas, entonces la superficie total equivale al **100,0 %** de los casos y en términos de probabilidad toma el valor de la **unidad.**



Asociar superficie a casos, porcentaje o probabilidad, es la base para el cálculo de probabilidades con variables numéricas continuas.

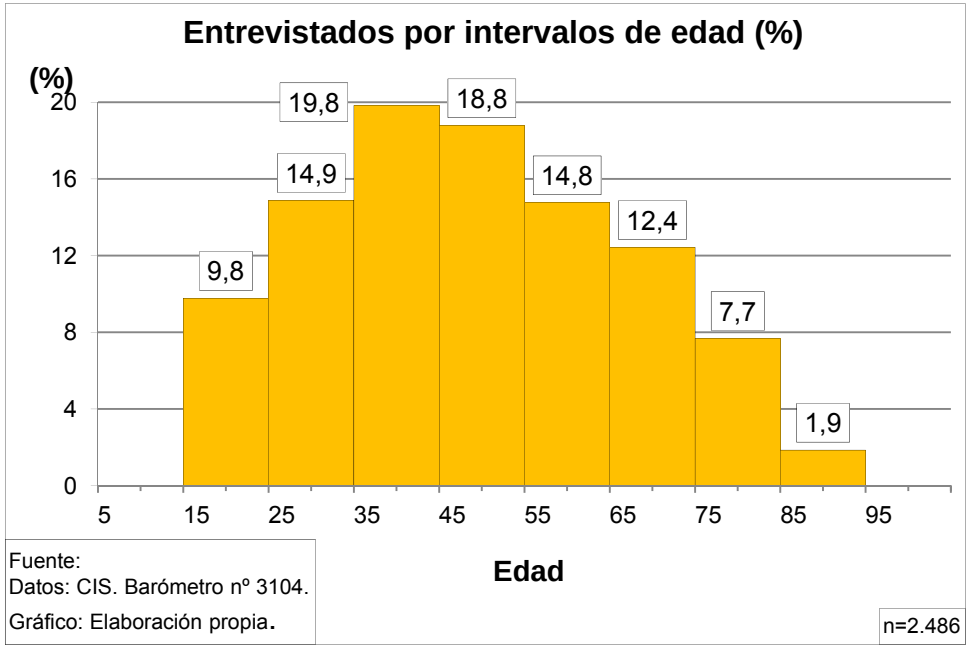
Histograma de frecuencias

Si se representan las frecuencias relativas, el aspecto y la relación es la misma. Lo que aportaría este tipo de gráfico es:

- Ver la diferencia de la frecuencia entre intervalos en puntos porcentuales
- Si la muestra fuese representativa daría una idea de magnitud en la población

En el histograma de frecuencias relativas, en cada columna, se representa el porcentaje de casos que hay por intervalo de edad y la lectura sería "**El 9,8% de los casos están en el intervalo de 15 a 25 años**". "**De 25 a 35 años están el 14,9% de los casos**", y así sucesivamente para todos los intervalos.

Como las columnas tienen igual ancho, si se considera que tiene valor unitario, al multiplicarlo por la altura, que es la frecuencia relativa, la **superficie** de cada columna representa su **frecuencia**. Por lo tanto, la **superficie total** del histograma representa el porcentaje total de casos, que es el **100%**, y en términos de probabilidad toma el valor de la **unidad**.

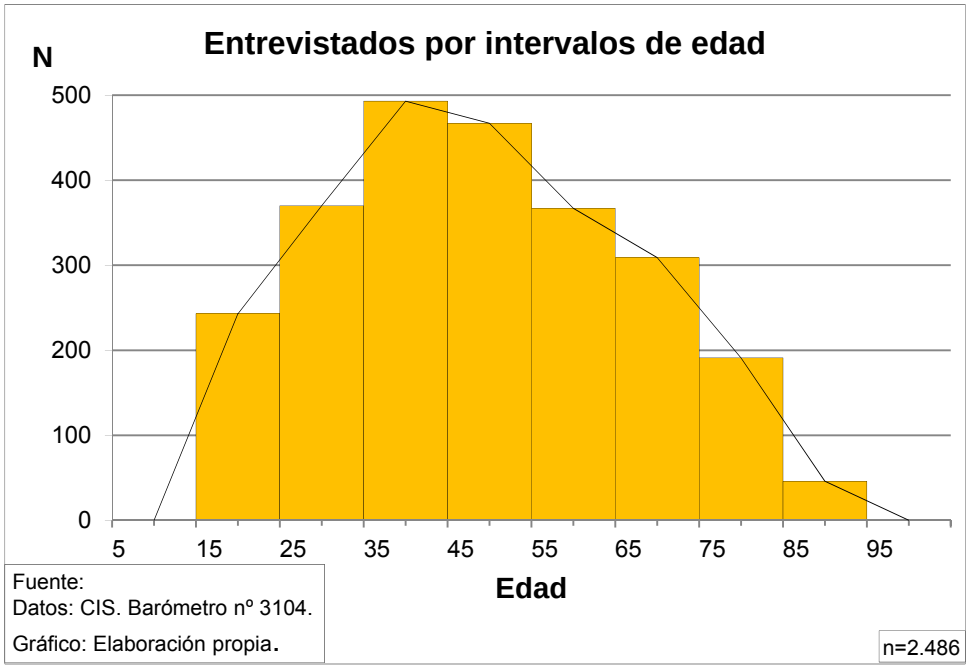


Polígono de frecuencias

Polígono de frecuencias

A partir del histograma, se construye el **polígono de frecuencias**. Se puede utilizar para el estudio, comprensión y descripción de las variables numéricas, conjuntamente con los estadísticos de **tendencia central**, **de dispersión** y **de forma**.

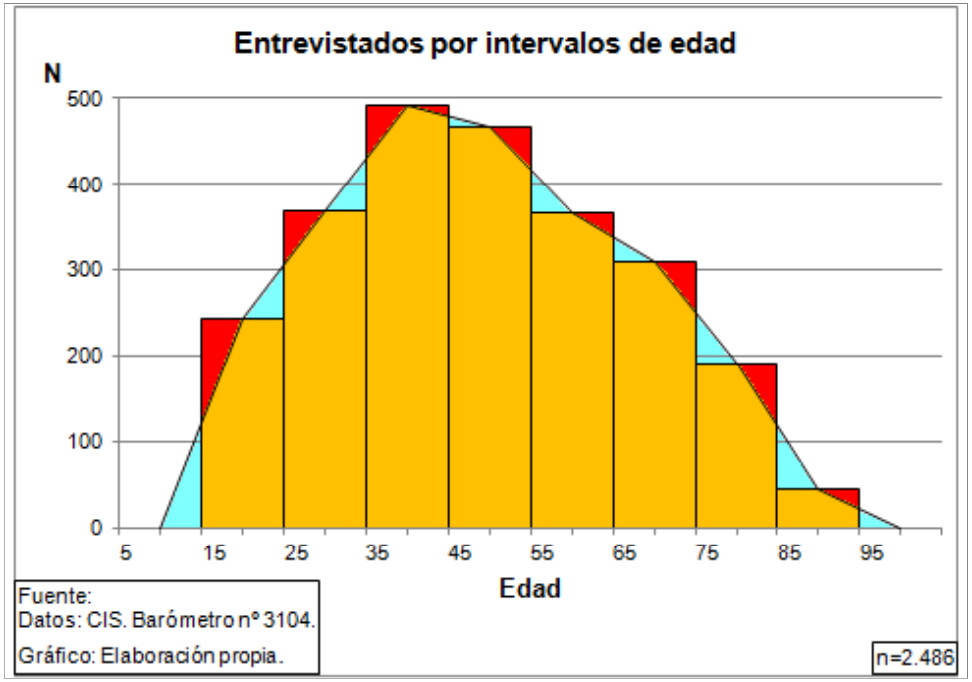
El polígono de frecuencias es la **línea quebrada** que resulta de **unir los puntos medios o marcas de clase de las caras opuestas a la base de cada intervalo o rectángulo**. El origen y final de la línea es en el eje, con valor de frecuencia cero, para cerrar la curva. Esta **línea quebrada**, por suavizado de los vértices, se considera que es una **curva**, que igual que en el polígono de frecuencias, se considera que la **superficie** contemplada por debajo de la curva así pintada y por encima del eje de abscisas, representa al **total de los casos**, que son el **100%** o que vale la **unidad** en términos de probabilidad



Polígono de frecuencias

Esta **igualdad** entre el **histograma** y el **polígono** se consigue porque la superficie que se incorpora de **fuera** del histograma (color azul) es igual a la superficie que se **omite** de las columnas (color rojo). Considerados de dos en dos, los triángulos opuestos por el vértice tienen la misma superficie por tener los lados y ángulos iguales.

En el polígono se pueden representar, mediante **vectores**, los estadísticos de **tendencia central**, **de forma** y **percentiles**.



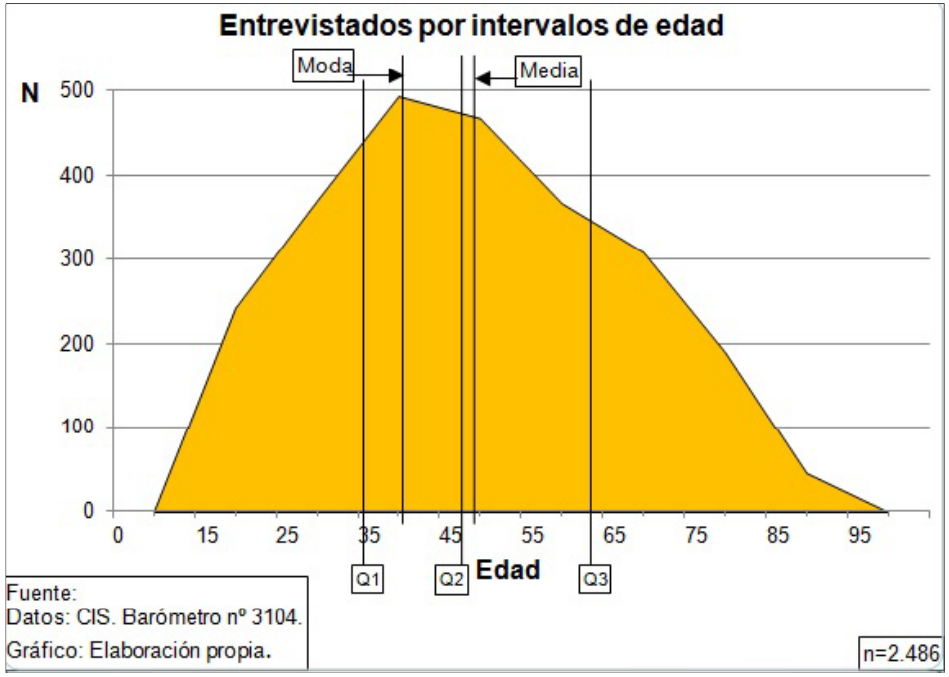
Polígono de frecuencias

En el polígono se pueden representar, mediante vectores, los estadísticos, y en este caso se han representado la **media**, la **moda**, el cuartil primero (**Q1**), el cuartil segundo (**Q2**) o mediana y el cuartil tercero (**Q3**).

La **media** (49,75 años) sería un centro de gravedad que indica que la suma de la diferencia de cada caso, inferior a la media, respecto a la media, es igual a la suma de los valores de la diferencia de cada caso, superior a la media, respecto a la media, pero con el signo cambiado. La suma es cero.

La **moda** (43 años) sería el valor que se repite más veces. Los percentiles son valores que definen superficies. Para los **cuartiles**, la superficie que hay por debajo del cuartil **Q1** representa al **25% de los casos**. La superficie que hay por debajo del cuartil **Q2** representa al **50% de los casos**, por lo tanto es la **mediana** y por debajo del cuartil **Q3**, está el **75% de los casos**. Gráficamente se puede ver que hay más dispersión en los extremos que en el centro, porque los recorridos o amplitudes son mayores.

Estadísticos		
Edad de la persona entrevistada		
Media		49,75
Mediana		48,50
Moda		43
Asimetría		0,20
Curtosis		-0,83
Percentiles	Q1, 25	36,00
	Q2, 50	48,50
	Q3, 75	64,00



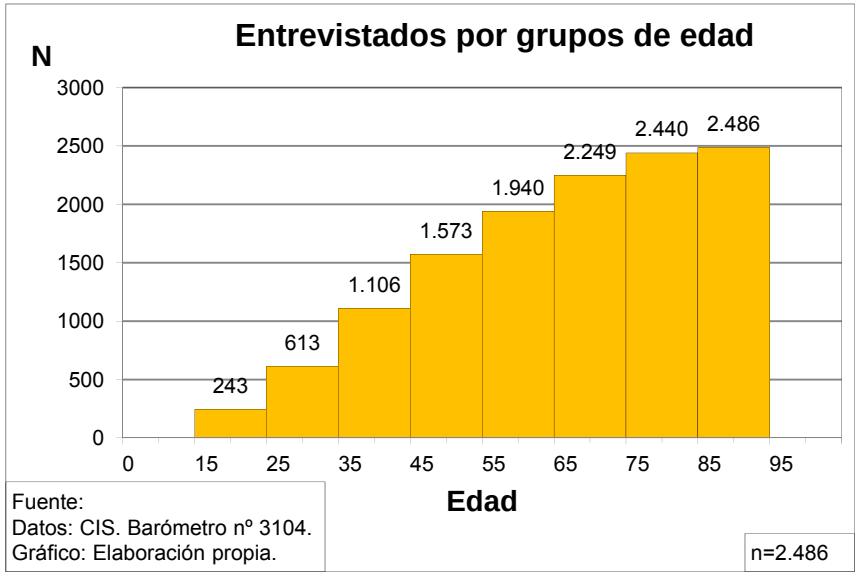
Histograma y polígono de frecuencias

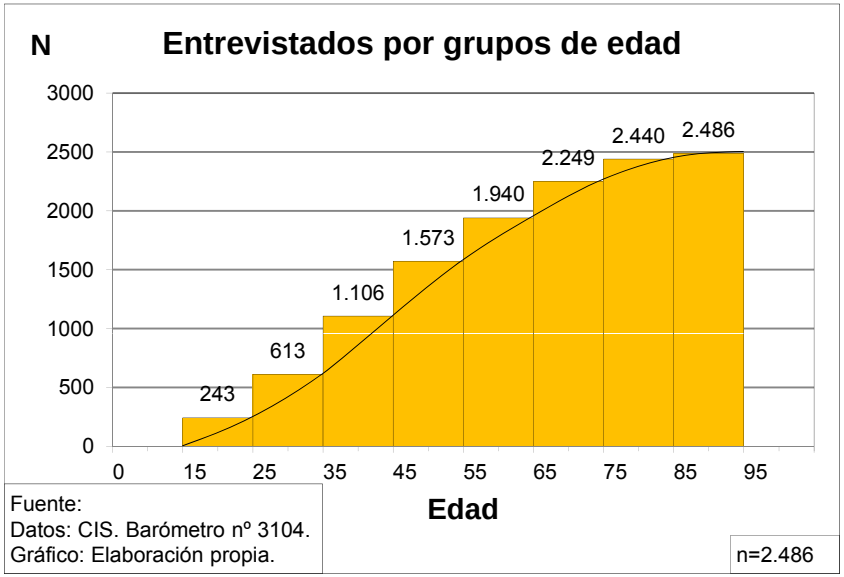
El histograma se puede presentar como histograma acumulado y a partir de éste el polígono de frecuencias acumuladas.

El histograma acumulado muestra, en cada columna, la frecuencia absoluta o relativa del intervalo en curso más la frecuencia de la anterior. La última columna acumula la frecuencia total.

Una lectura: "**Con menos de 55 años hay 1.573 casos**".

El polígono de frecuencias acumulado es la curva del histograma acumulado y se traza uniendo el vértice de frecuencia cero de la primera columna con el vértice opuesto que indica la frecuencia del intervalo, y así sucesivamente. El último punto de la curva muestra la frecuencia total.



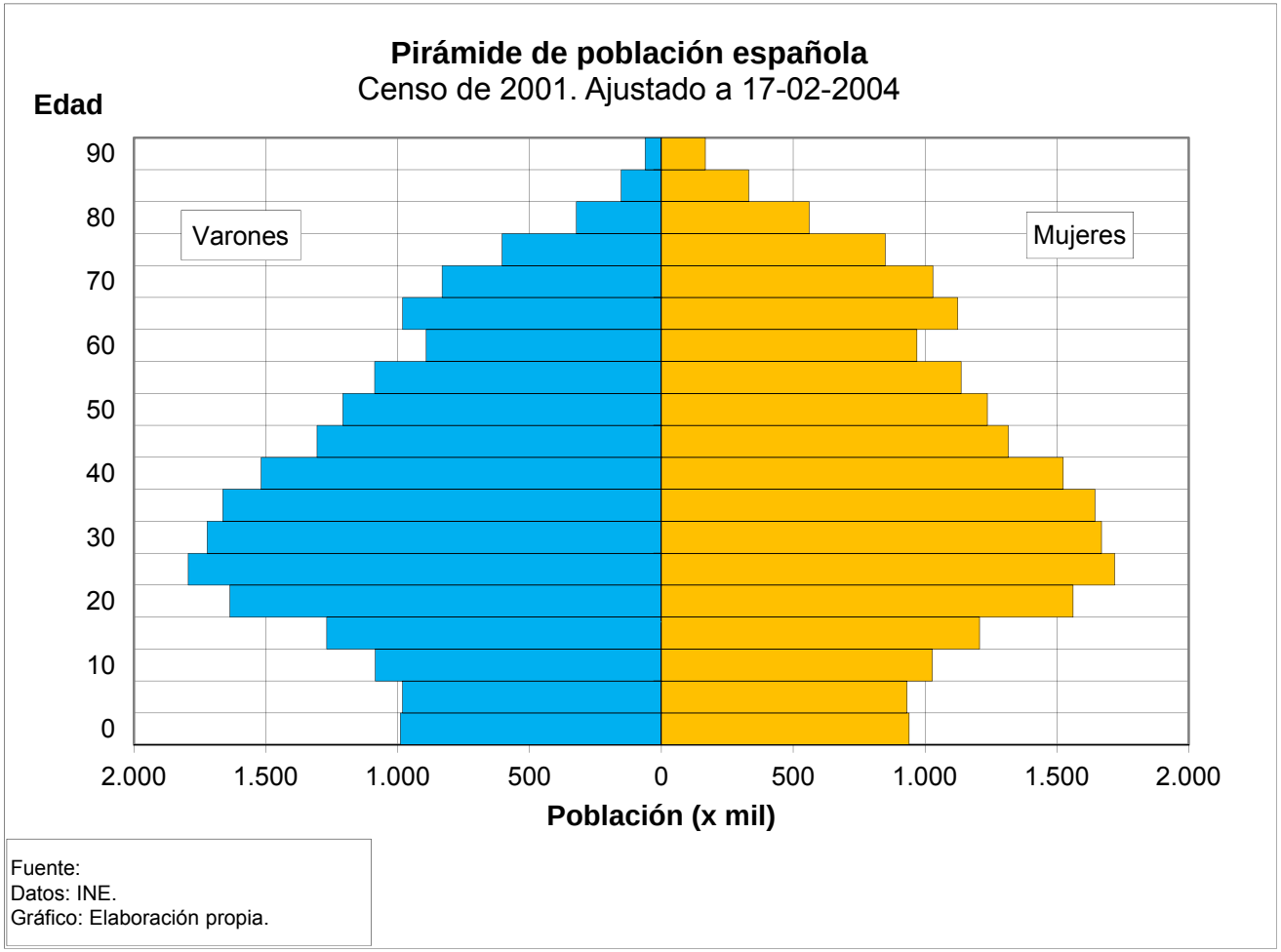


Pirámide

Una aplicación del histograma en formato de barras sería la **pirámide de población**.

El eje *X* puede estar en valores absolutos o porcentajes, si es la población, si no, sólo en porcentajes

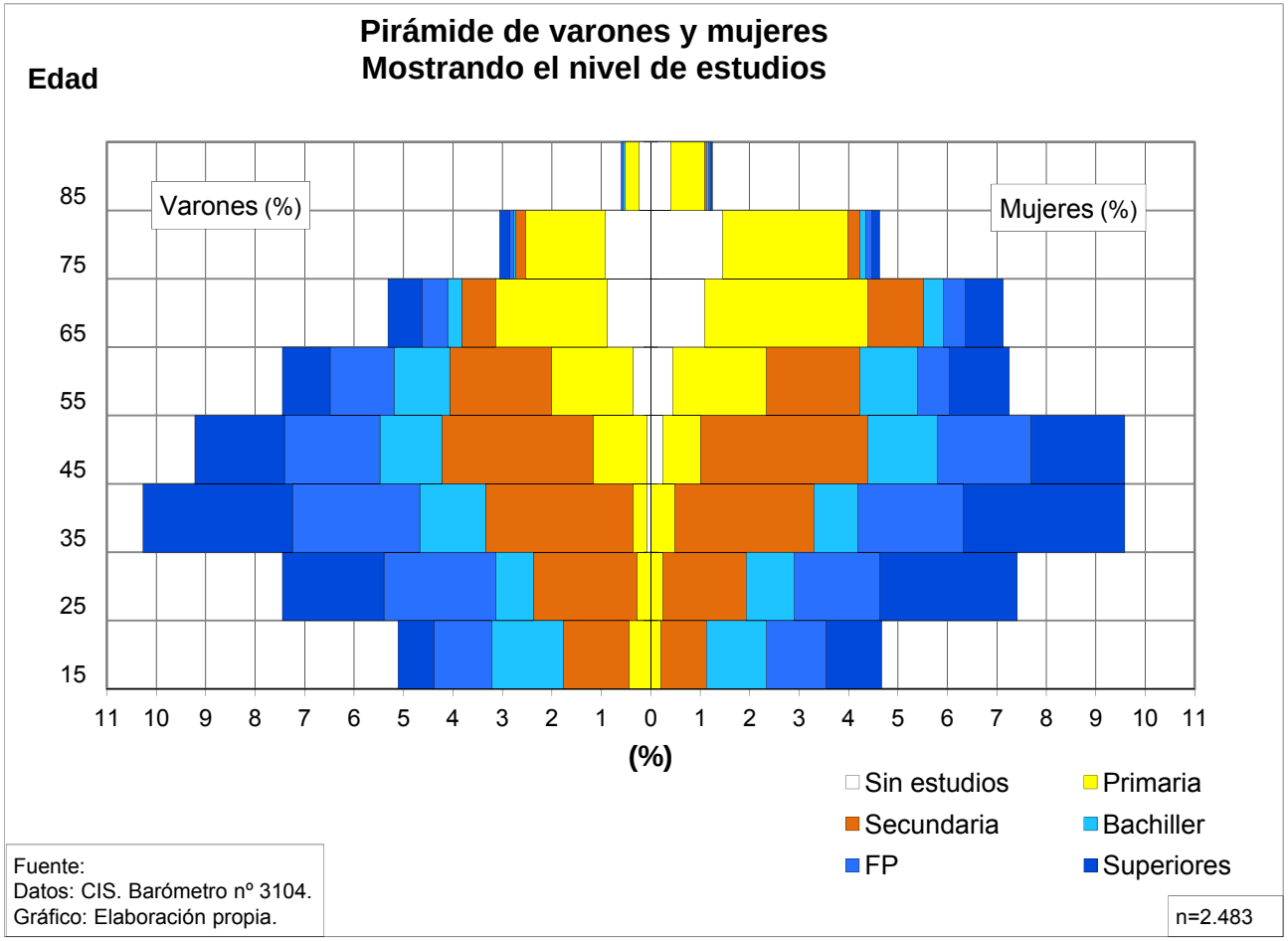
Para ver: Ejemplo de pirámide
(Statistics Canada, 2015)



Fuente:
Datos: INE.
Gráfico: Elaboración propia.

Pirámide

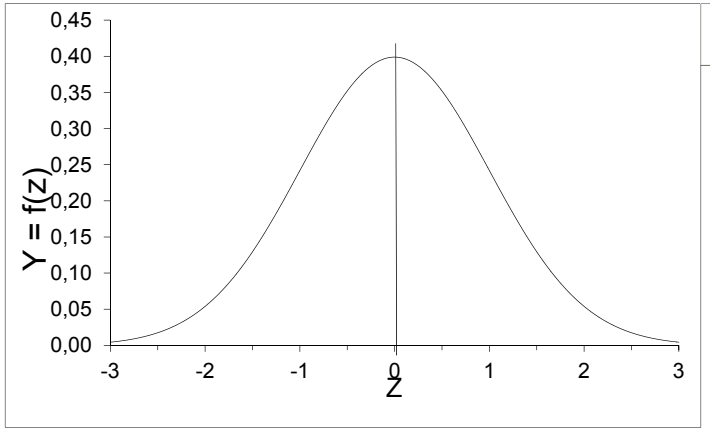
En esta pirámide se muestra la composición de **nivel de estudios** por **intervalo de edad**.



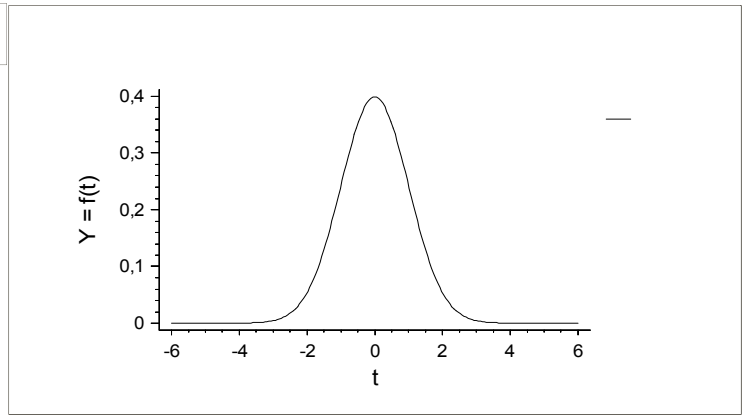
Variables estandarizadas

Aplicaciones o variaciones del polígono de frecuencias se pueden considerar las distribuciones de probabilidad y las más utilizadas en Ciencias Sociales son: la ***Z***, la ***Chi-cuadrado*** (χ^2), la ***t*** y la ***F*** (De la Puente, 2015).

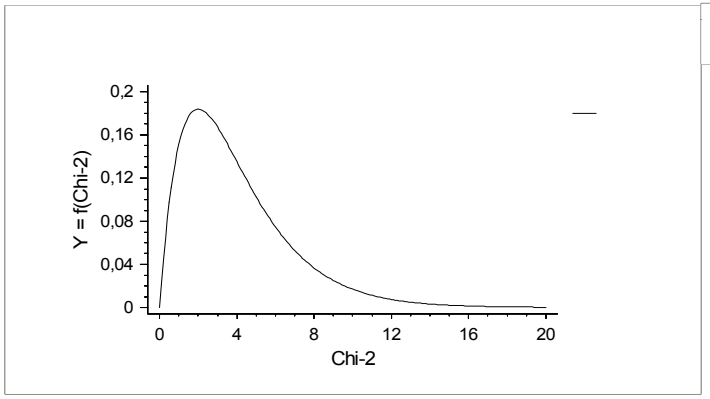
Z



t



χ^2



F

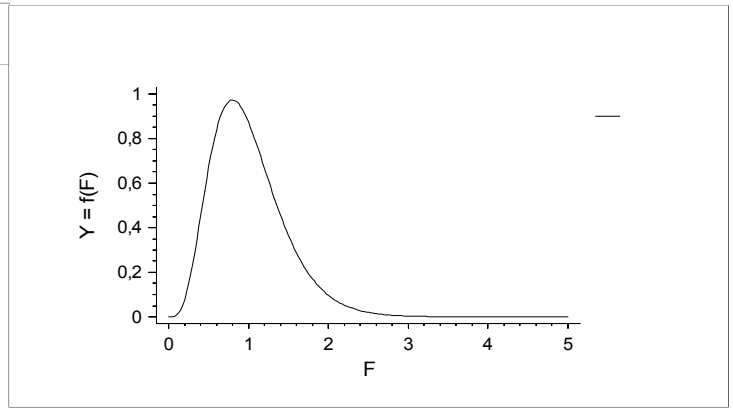


Gráfico de dispersión

Gráfico de dispersión

El **gráfico de dispersión** o gráfico **de punto a punto**, considerando el caso de dos variables, es la representación gráfica de variables numéricas, o consideradas numéricas, y ayuda al estudio de la **asociación** (Covarianza y correlación r de Pearson) y la **regresión lineal** (Ajuste de una recta por el método de mínimos cuadrados ordinarios) entre las dos variables.

El gráfico es la representación de las dos variables en un sistema de coordenadas cartesianas de dos dimensiones y si los valores son positivos, sólo se utiliza el cuadrante-I.

En el **eje horizontal** de abscisas o *X* se representa una variable, normalmente la independiente o considerada la independiente. En el **eje vertical** de ordenadas o *Y* se representa la variable considerada dependiente. Si hay dos variables independientes, el gráfico es de tres dimensiones y si hay **n** variables independientes el gráfico es de **n+1** dimensiones.

Así dispuesto, cada caso o unidad de observación se representa en el plano por un punto, según el valor o coordenada en la variable del eje ***X*** y el valor o coordenada en la variable del eje ***Y*** .

En la **asociación lineal** no hay relación de dependencia entre las variables, pero como en la **regresión lineal** sí, se asume relación de dependencia y como la asociación es la antesala de la regresión, siempre se considera la **relación entre las variables** para su presentación en los ejes.

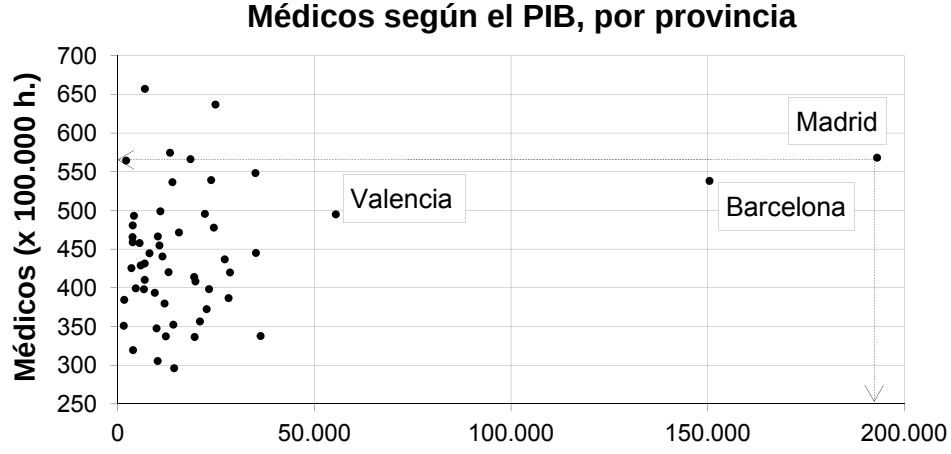
Los datos que se van a utilizar para representar en este gráfico son el **PIB por habitante** y el **número de médicos por cada 100.000 habitantes**, ambos del año 2008, por provincia y por comunidad autónoma, en España (Fuente: INE Indicadores sociales 2010).

Gráfico de dispersión

En este caso, si existe relación, se va a considerar que el **número de médicos** es dependiente del **PIB**. Por lo tanto, en el **eje *X*** se va a presentar el **PIB** como variable **independiente**, y en el eje ***Y*** el **número de médicos**, que se va a considerar o proponer como **dependiente** del PIB o que es función del PIB. Las unidades de los ejes están escaladas de tal manera que el segmento unitario permita representar desde el valor mínimo al máximo de cada una de las variables y se pueda representar dentro de la hoja o marco del gráfico.

El origen de los ejes es el punto cero, que es en donde se cortan los dos ejes. Pero cuando la distancia entre el punto cero de un eje y el valor mínimo de la variable que se representa en ese eje es grande, el espacio blanco que queda y que reduce la visibilidad del gráfico, se suprime y entonces el origen se representa como un valor distinto de cero. Como en este caso.

Cada **punto** del gráfico es una **provincia** y sus **coordenadas** (x, y) son el **PIB por habitante** y el **número de médicos por 100.000 habitantes**.
Ejemplo: las coordenadas de Madrid son (193.050, 568).



Fuente:
Datos: INE. Indicadores Sociales 2010.
Gráfico: Elaboración propia.

Gráfico de dispersión

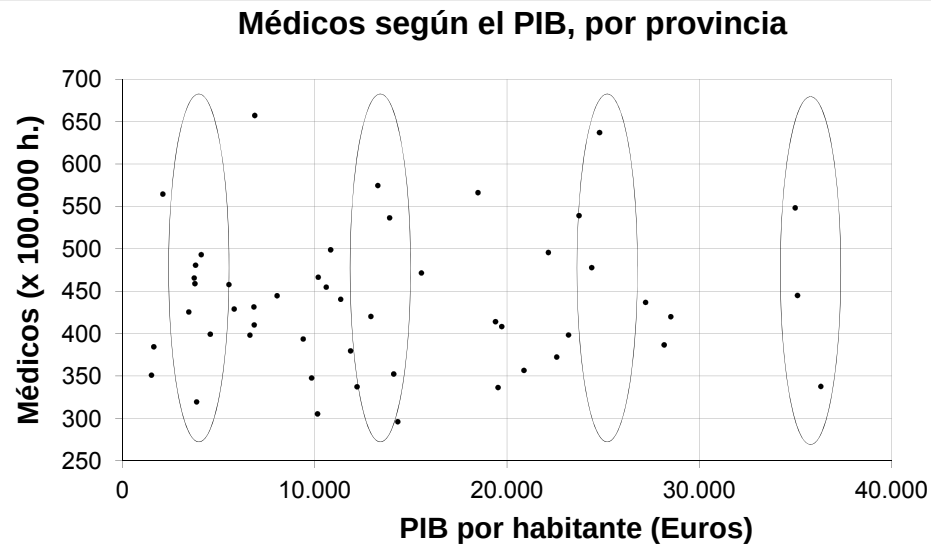
Como el **PIB** de **Barcelona**, **Madrid** y **Valencia** es mayor que el resto de las provincias, hace que estas se vean concentradas en los valores bajos del eje.

Para ver la relación más detallada, se pueden **suprimir** las provincias de renta más alta, y se consigue un efecto zoom.

El gráfico de punto a punto o dispersión ayuda a la interpretación y análisis de la asociación y regresión lineal. En este caso y sin utilizar los valores de los estadísticos, se puede utilizar el **sentido común** documentado en base a la experiencia de otros casos anteriores.

El gráfico indica que **no hay asociación** o que la asociación es dispersa. Esto es indicativo de que no hay relación lineal entre el **PIB** y el **número de médicos**. Esto es, que el número de médicos no depende del PIB por habitante de la provincia.

El gráfico indicaría que la salud no está en función del PIB de la provincia, ya que a cualquier PIB le corresponde cualquier número de médicos.



Fuente:
Datos: INE. Indicadores Sociales 2010.
Gráfico: Elaboración propia.

Sin:
Barcelona, Madrid y Valencia

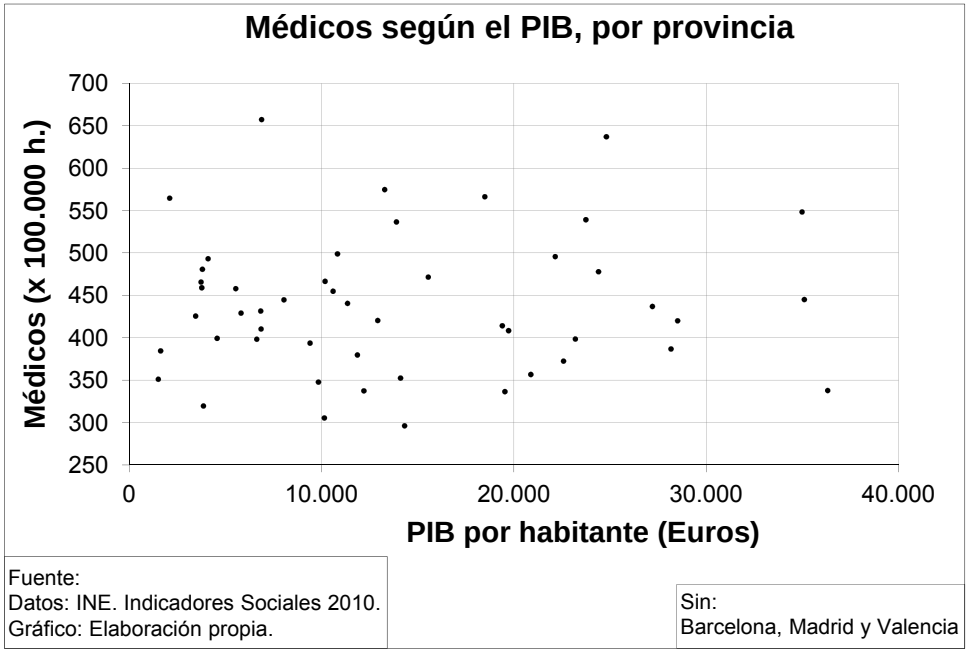
Gráfico de dispersión

Pero puede haber **otra interpretación**, y es que independientemente del **PIB** de la provincia, el **número de médicos** oscila en el rango de **menos de 300** médicos (Almería: 296,13 médicos/100.000 h. PIB: 14.311,50 euros/h.) hasta **más de 650** médicos (Salamanca: 657,30 médicos/100.000 h. PIB: 6.876,40 euros/h.).

Hasta aquí se puede considerar la labor de un **técnico** y ahora se aplicaría el conocimiento sobre la situación, para interpretar el gráfico y hacer las aplicaciones correspondientes. Una pregunta puede ser **¿Qué produce la diferencia de médicos?**

Por lo tanto, la **utilidad** de estos gráficos es independiente de que haya o no asociación y relación lineal.

El **siguiente paso** podría ser comprobar si esta dispersión se neutraliza utilizando las **comunidades autónomas**.



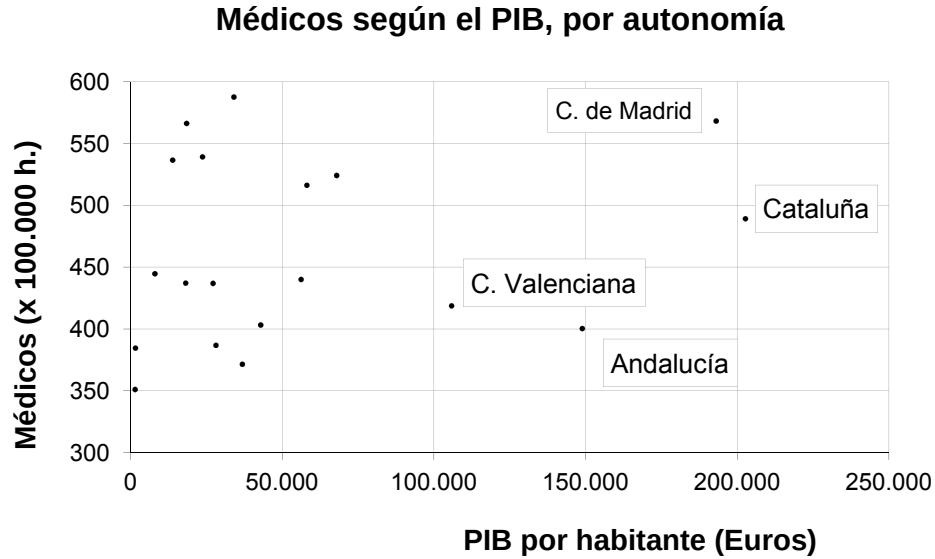
Fuente:
Datos: INE. Indicadores Sociales 2010.
Gráfico: Elaboración propia.

Sin:
Barcelona, Madrid y Valencia

Gráfico de dispersión

Al comprobar si esta dispersión se neutraliza utilizando las comunidades autónomas, se observa que se mantiene.

Se pueden buscar **otros criterios** para ver qué es lo que puede neutralizar la dispersión, o qué la explicaría.



Fuente:
Datos: INE. Indicadores Sociales 2010.
Gráfico: Elaboración propia.

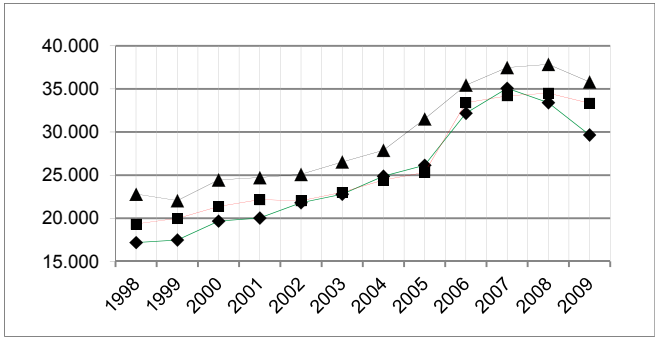
Series temporales

Una **serie temporal** es una variable en la que sus **datos** están **ordenados** en base a una **secuencia temporal** o **cronológica** que está definida por otra variable, medida u observada en intervalos de tiempo siempre iguales.

La representación gráfica de una serie temporal se realiza en un sistema de coordenadas cartesianas de dos dimensiones. En el **eje horizontal** o *X* se representa la **variable *tiempo,*** y en el **eje vertical** o *Y* se representa la variable de la **serie**.

En el eje *Y* se pueden representar varias variables o series temporales. Es recomendable que las series temporales estén en la **misma unidad de medida**. Se pueden utilizar dos unidades de medida, estando la segunda representada en el eje ***Y* secundario** (a la derecha del gráfico). La unidad de tiempo debe ser la misma para todas las series.

En el gráfico, se representan por puntos los valores que toma la serie en cada una de las unidades de tiempo. Planteado de esta manera, es un **gráfico de dispersión** o **punto a punto**. Después, se unen con una línea todos los puntos y se configura lo que se llama la "**serie temporal**".



Series temporales

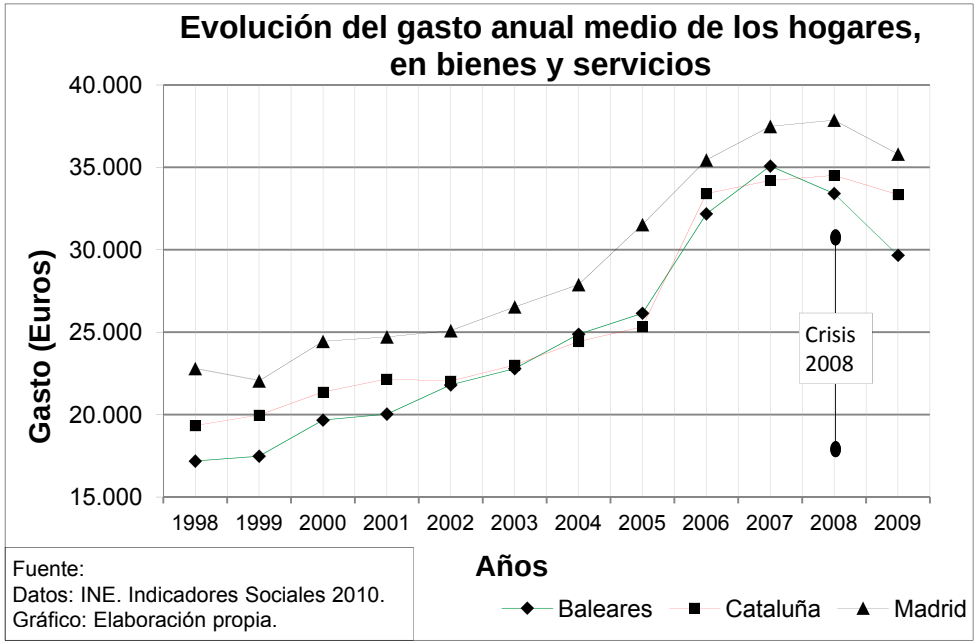
En este ejemplo se representan las series de **"Gasto anual medio de los hogares en bienes y servicios"** entre los años 1998 y 2009. Los datos se han extraído de los Indicadores Sociales 2010 del INE.

Se representan sólo las series de las comunidades de las **Islas Baleares**, **Cataluña** y **Madrid**, por cuestiones de brevedad y claridad en la exposición.

Además de la evolución en el tiempo del gasto en bienes y servicios, de cada una de las comunidades, se puede comparar el gasto y la evolución de las tres comunidades.

Una lectura: **Las series tienen una tendencia creciente hasta 2007. En 2008 se produce la estabilización o inicio de la caída y desde 2008, año de la crisis, se produce la caída de las tres series**.

El gráfico puede ir acompañado con etiquetas identificativas con los hechos que pueden haber influido en la evolución de las series.



Fuente:
Datos: INE. Indicadores Sociales 2010.
Gráfico: Elaboración propia.

Series temporales

Ejemplo de gráfico de series temporales con dos ejes *Y*. Evolución del IBEX 35 (Eje *Y*) y de su capitalización (Eje *Y* secundario) desde 1992 a 2011 (Eje *X*). Fuente: (Fernández, Aguirreamalloa, & Corres, 2011) (Añadidos los cuadros de texto).

Figura 1

Evolución del IBEX 35 y de su capitalización (millones de euros)

Series temporales

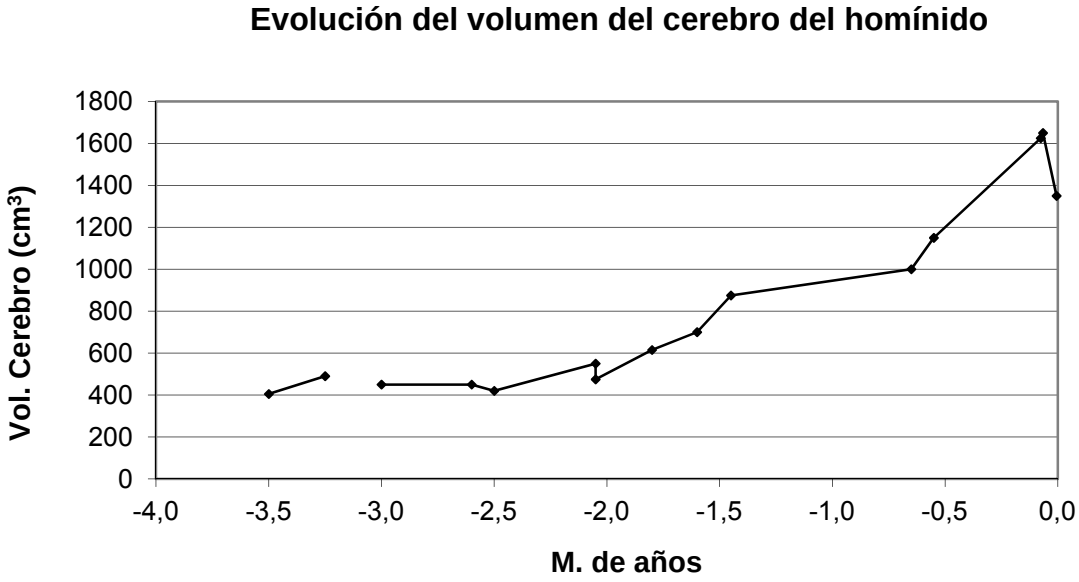
Ejemplo de gráfico de series temporales con anotaciones. Fuente: (De la Quintana, 2015).

Series temporales

Ejemplo: Serie temporal y análisis de regresión. Fuente: (De la Puente, 2011).

Primera expansión del volumen del cerebro del *Homo* (2 Millones de años): b=630,25; β = 0,98; r^2 = 0,96; sig.). Por cada millón de años el volumen del cerebro se incrementa en más de 600 cm^3.

Segunda expansión (500.000 años): b = 1.018,23; β = 1,00; r^2 = 1,00; sig.). Por cada millón de años el volumen del cerebro se incrementa en 1.000 cm^3.



La serie indica el desarrollo del volumen del cerebro. La regresión, la relación entre el volumen y el tiempo. Podría interpretarse que el paso del tiempo es la causa. Pero las causas reales tienen diferentes interpretaciones.

Bibliografía

Bibliografía

De la Puente, C. (2011). *Fundamentos de Neurosociología.* Madrid: Complutense.

De la Puente, C. (April de 2014b). Proposal for a reasonable model of the visual system. Principles of Clinical Neurosociology. *Sociology Study, 4*(4 (35)), 360-383.

De la Puente, C. (2015d). *Estadística descriptiva e inferencial y una introducción al método científico. Con un apéndice al método.* Madrid: IDT CB.

Fernández, P., Aguirreamalloa, J., & Corres, L. (2011). *IBEX 35: 1991-2010. Rentabilidad y creación de valor.* IESE Business School-Universidad de Navarra: http://www.iese.edu/research/pdfs/DI-0890.pdf

Galton, F. (1889). *Natural inheritance.* London: Macmillan.

Graunt, J. (25 de January de 1662/1996). *Natural and political observations mentioned in a following index, and made upon the bills of mortality.* edstephan: http://www.edstephan.org/Graunt/bills.html

Herrera Arellano, J. L. (s.f.). *Símbolos*. jorgeluisherrera: http://jorgeluisherrera.com/simbolos.html

Muñoz, A. M., & Molero, J. A. (julio-septiembre de 2012). Altamira, un legado del paleolítico español. *Revista de creación literaria y humanidades, 2*(77), 12.

Pearson, K. (1895). Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous. *Philosophical Transactions of the Royal Society of London A, 186*, 343-414.

Pedro1267. (25 de noviembre de 2015). *Bosque*. Wikipedia: https://es.wikipedia.org/wiki/Bosque

Spence, I. (2006). William Playfair and the Psychology of Graphs. *ASA Section on Statistical Graphics*, 2426-2436.

Statistics Canada. (23 de November de 2015). *Historical age pyramid*. Statistics Canada: https://www12.statcan.gc.ca/census-recensement/2011/dp-pd/pyramid-pyramide/his/index-eng.cfm

Wikipedia. (16 de noviembre de 2015). *Numeración griega*. Wikipedia: https://es.wikipedia.org/wiki/Numeraci%C3%B3n_griega

Wikipedia. (2 de August de 2015a). *Chart*. Wikipedia: https://commons.wikimedia.org/wiki/Chart

Wikipedia. (23 de August de 2015b). *Bar chart*. Wikipedia: https://commons.wikimedia.org/wiki/Bar_chart#History

Wikipedia. (31 de August de 2015c). *William Playfair*. Wikipedia: https://en.wikipedia.org/wiki/William_Playfair

Wikipedia. (10 de October de 2015d). *Scatter plot*. Wikipedia: https://en.wikipedia.org/wiki/Scatter_plot

WordPress.com. (28 de marzo de 2014). *Grafística. La Historia de la Escritura. Capítulo I: Mesopotamia*. Periciales Lázaro: https://pericialeslazaro.files.wordpress.com/2014/03/escritura-cuneiforme.jpg.

Estadística y Gráficos
Statistic and graphics

Diagrama de barras, Histograma,
Polígono de frecuencias y Gráfico de dispersión

Bar chart , Histogram, Frequencies polygon & Scatter plot

Muchas gracias

Carlos DE LA PUENTE VIEDMA
Sociología IV – UCM