Estadística aplicada a la comunicación

Tema 5: Análisis de datos cuantitativos I: estadística descriptiva

a. Análisis univariante

OpenCourseWare UPV/EHU

Unai Martín Roncero
Departamento de Sociología 2
unai.martin@ehu.eus

Índice

Antes de empezar …

Ω= colectivo de estudio o población

N= Tamaño del colectivo n= Tamaño de la muestra

ω_i= cada uno de los elementos del colectivo, i= 1, ….N

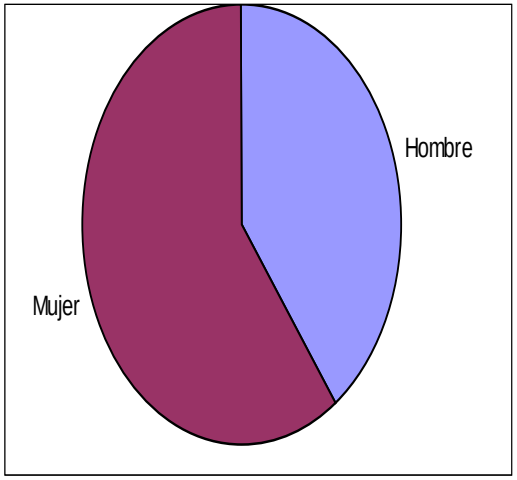
X, Y…= cada variable objeto de estudio

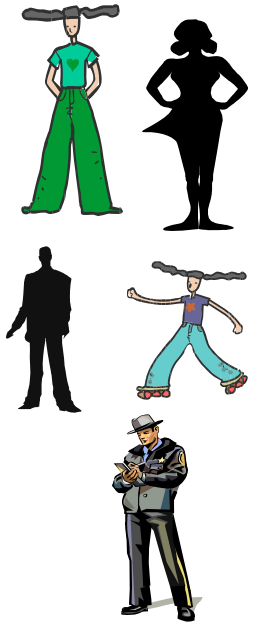
$x_1,\ldots,x_k$ = conjunto de valores que toma la variable X en el colectivo. Por tanto, k es el número de valores distintos que toma la variable.

Distribución de frecuencias y representaciones gráficas

Tablas de frecuencias y gráficos: son formas equivalentes de recoger de una manera **rápida** la información de la matriz de datos de una forma **ordenada y resumida.**

Muy útil en presentación de resultados y análisis exploratorio





Sexo	n
Hombres	2
Mujeres	3

Distribución de frecuencias y representaciones gráficas

Tablas de frecuencia: resumen los datos de una variable, perdiendo, en algunos casos, algo de información.

Nivel de estudios	F. Absoluta	F. Relativa	F. Rel acumulada	F. Abs acumulada
Primarios	317	63,4	63,4	317
Secundarios	132	26,4	89,8	449
Superiores	51	10,2	100	500
Total	500	100		

Distribución de frecuencias y representaciones gráficas

Tablas de frecuencia, contienen básicamente para cada clase c_i:

-**Frecuencia absoluta** (n_i), número de veces que se repite ese valor (x_i) o modalidad en el total de individuos.

-**Frecuencia relativa** (f_i), número de veces, en **tantos por uno** que se repite ese valor (xi) o modalidad en el total de individuos.

$$f_i = \frac{n_i}{n}$$

También se utiliza el porcentaje $p_i = f_i * 100$

- **Frecuencia absoluta acumulada** (N_i), suma de n_i acumuladas hasta esa clase

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{j=1}^{i} n_j$$

<u>No variables nominales</u>

- **Frecuencia relativa acumulada** (F_i)

$$F_i = \frac{N_i}{n} = \frac{n_1 + \ldots + n_i}{n} = f_1 + \ldots + f_i = \sum_{j=1}^{i} f_j$$

(Barón 2004) (Barcena y otras 2003)

Distribución de frecuencias y representaciones gráficas

Modali.	Frec. Abs.	Frec. Rel.	Frec. Abs. Acumu.	Frec. Rel. Acumu.
C	n_i	f_i	N_i	F_i
c_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = \frac{N_1}{n} = f_1$
...	...	...	...	...
c_j	n_j	$f_j = \frac{n_j}{n}$	$N_j = n_1 + \ldots + n_j$	$F_j = \frac{N_j}{n} = f_1 + \ldots + f_j$
...	...	...	...	...
c_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
	n	1		

(Barón 2004)

Distribución de frecuencias y representaciones gráficas

Población de 10 y más años no estudiante por nivel de instrucción en la CAPV. 2001

	n_i	N_i	f_i	F_i	p_i	P_i
Analfabetos	13027	13027	0,008	0,008	0,80	0,80
Sin estudios	80802	93829	0,049	0,057	4,93	5,73
Primarios	775960	869789	0,474	0,531	47,38	53,10
Profesionales	234793	1104582	0,143	0,674	14,34	67,44
Secundarios	236046	1340628	0,144	0,819	14,41	81,85
Medio-superiores	112871	1453499	0,069	0,887	6,89	88,74
Superiores	184400	1637899	0,113	1,000	11,26	100,00
Total	1637899		1		100,00	

Fuente: Eustat

Distribución de frecuencias y representaciones gráficas

Categorías de las variables deben ser:

✓Exhaustivas: todos los individuos deben poder ubicarse en una categoría. Recogen todas las opciones posibles

✓Excluyentes: todos los individuos deben colocarse en sólo una categoría, no en más de una. No deben solaparse entre sí.

Distribución de frecuencias y representaciones gráficas

Representaciones gráficas, complementan a las distribuciones de frecuencias a la hora de obtener una visión rápida de la información.

Los gráficos deben:

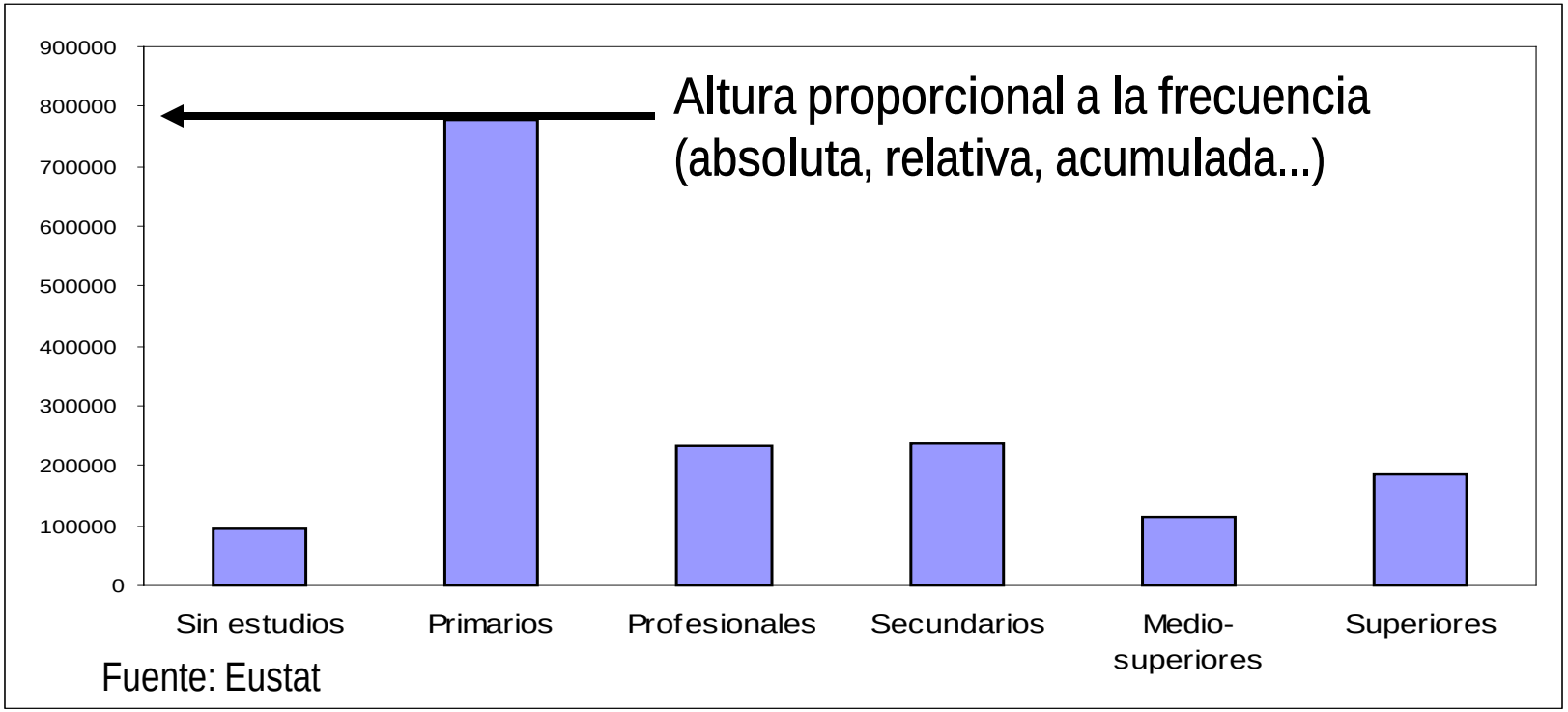
- Ayudar a obtener una visión fidedigna de la(s) variable(s)
- Ser claros, tener título y fuente
- Deben captar la atención pero ser sencillos, claros y precisos

El tipo de variable condiciona el tipo de gráfico

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras

Gráfico I "Población de 10 y más años no estudiante por nivel de instrucción en la CAPV. 2001"



Fuente: Eustat

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

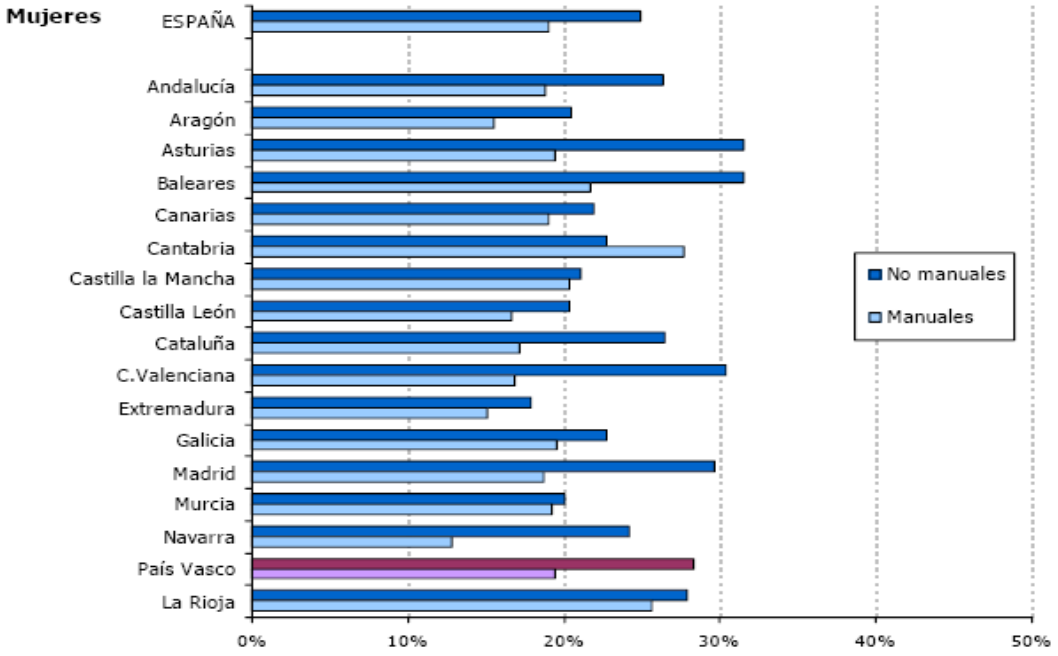
- Variables cualitativas y discretas
- Comparar magnitudes de cada modalidad
- En el eje vertical se representan las modalidades y en el horizontal su frecuencia (absoluta o relativa)
- El orden de las modalidades en el gráfico puede ser según su orden “natural” (ordinales), alfabético (útil cuando hay muchas modalidades), según la magnitud (de menor a mayor) o aleatorio.

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

Las barras pueden ser horizontales (muchas categorías, nombres largos etc.)

Distribución de las prevalencias de visitas al dentista según la clase social en mujeres, en las CC.AA. de Estado Español el año 2003.



Fuente: VV.AA

Distribución de frecuencias y representaciones gráficas

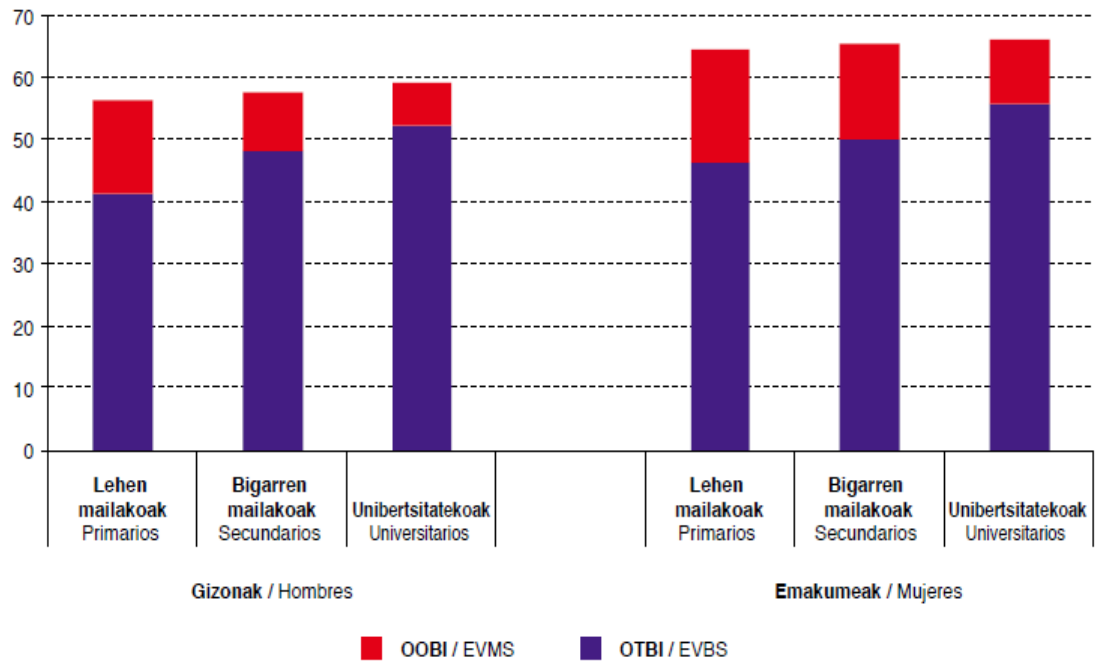
Diagrama/gráfico de barras:

Barras apiladas: permiten describir las diferencias entre cada modalidad o periodo en la magnitud general y desagregada

Esperanza de vida (total barra), esperanza de vida en buena salud (EVBS), esperanza de vida en mala salud (EVMS) a los 20 años según nivel de estudios y sexo en la C.A. de Euskadi. 1996-2001

22

Bizi-itxaropena (barra osoa), osasun onean bizitzeko itxaropena (OOBI), osasun txarrean bizitzeko itxaropena (OTBI) 20 urte izatean, ikasketa-mailaren eta sexuaren arabera, Euskal AEn. 1996-2001



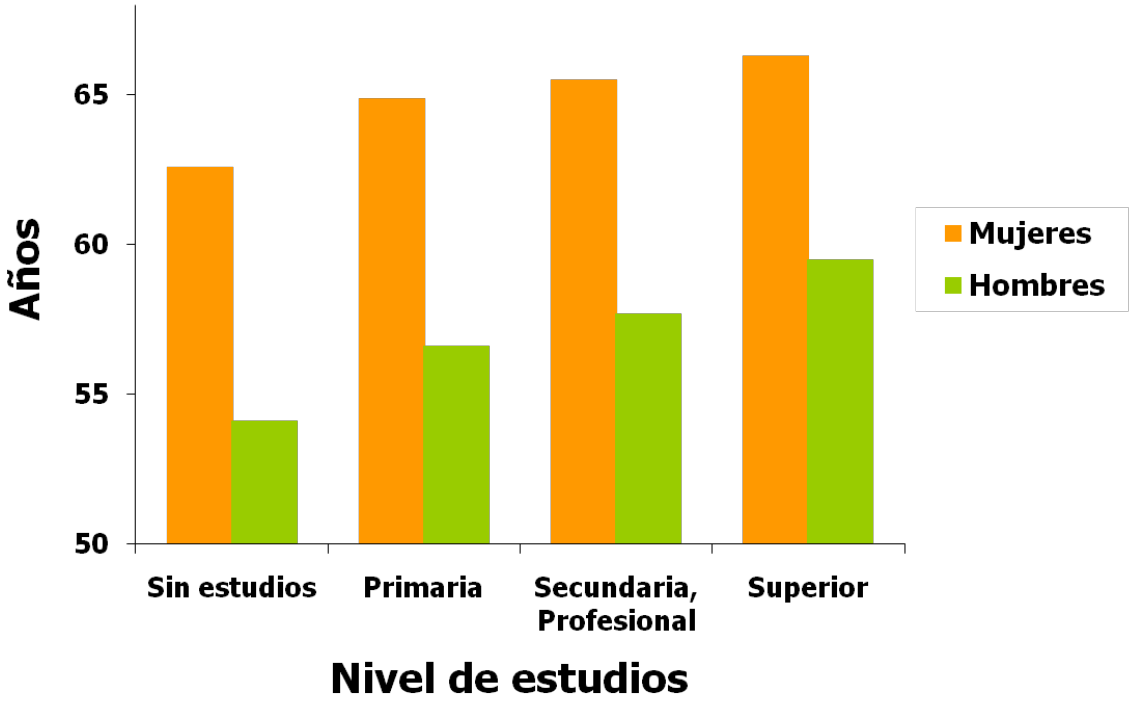
Iturria: Martín, 2011, EAEOlan eta Osasun Sailean oinarrituta, 2005.
Fuente: Martín, 2011 a partir de ESCAV y Departamento de Sanidad, 2005.

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

Barras agrupadas: permiten comparar las diferencias entre las modalidades o periodos y si esas diferencias varían en los grupos formadas según otra variable (por ejemplo sexo)

Esperanza de vida a los 20 años según el nivel de estudios, CAPV 1996-2001



Fuente: Servicio de Estudios e Investigación Sanitaria. Departamento de Sanidad

Distribución de frecuencias y representaciones gráficas

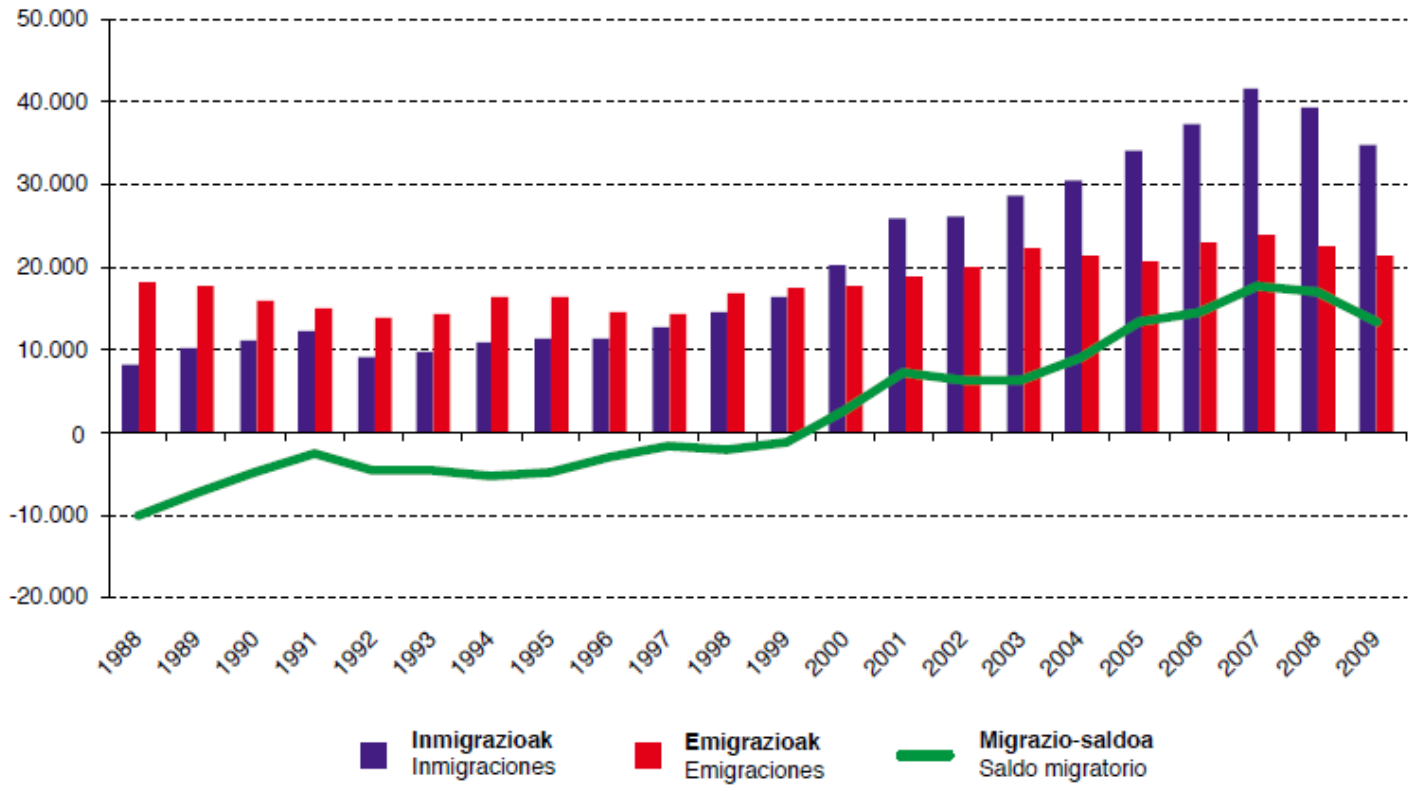
Diagrama/gráfico de barras:

Barras agrupadas:

Número de inmigrantes y emigrantes y saldo migratorio por año en la C.A. de Euskadi

23

Etorkinen eta emigranteen kopurua eta migrazio-saldoa, urtearen arabera, Euskal AEn



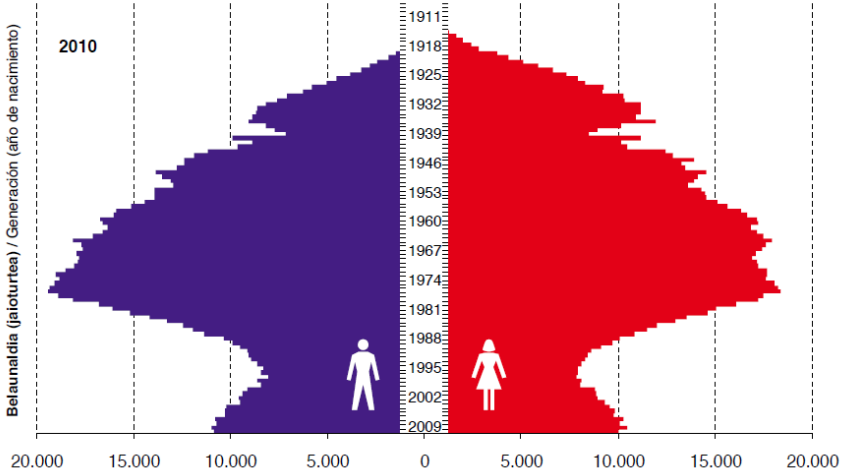
Iturria: Geuk egina, Eustaten oinarrituta.
Fuente: Elaboración propia a partir de Eustat.

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

Pirámides de población

Estructura por edad (año de nacimiento) y sexo de la C. A. de Euskadi **2-3** ***Adinaren (jaiotza urtea) eta sexuaren araberako egitura Euskal AEn***



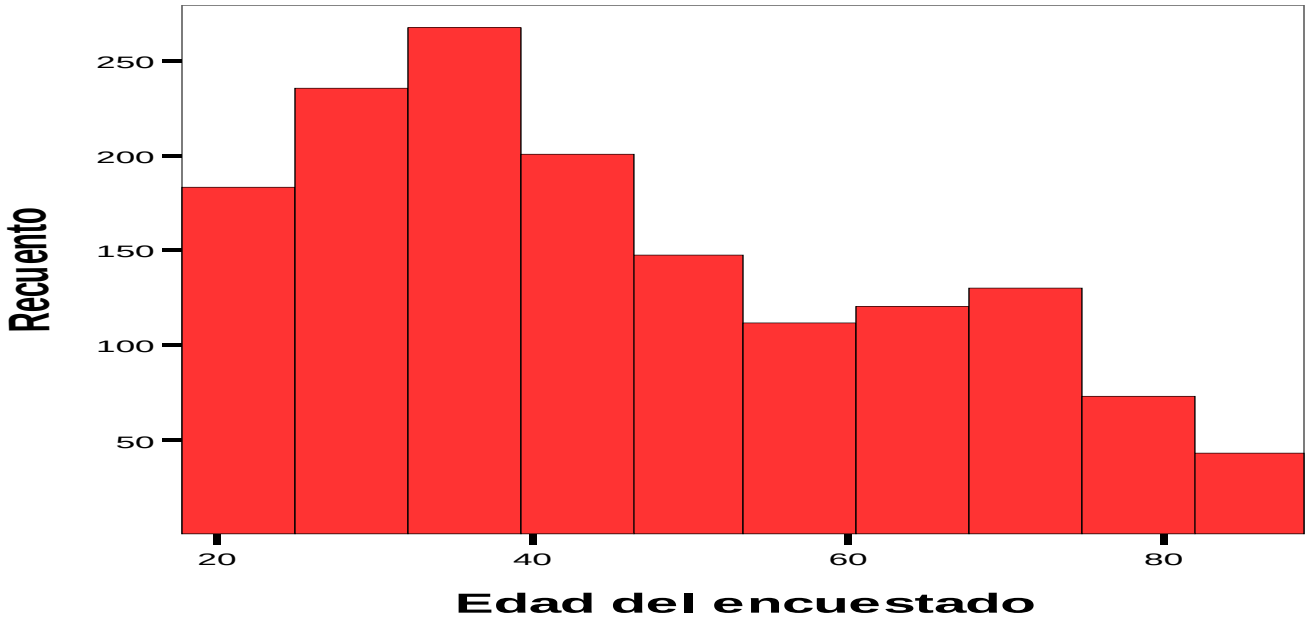
Iturria: Geuk egina, Eustaten oinarrituta.
Fuente: Elaboración propia a partir de Eustat.

Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

Histograma: Gráfico para variables continuas

Similar al diagrama de barras pero sin hueco, para dar idea de continuidad.
También puede tener un polígono de frecuencias asociado



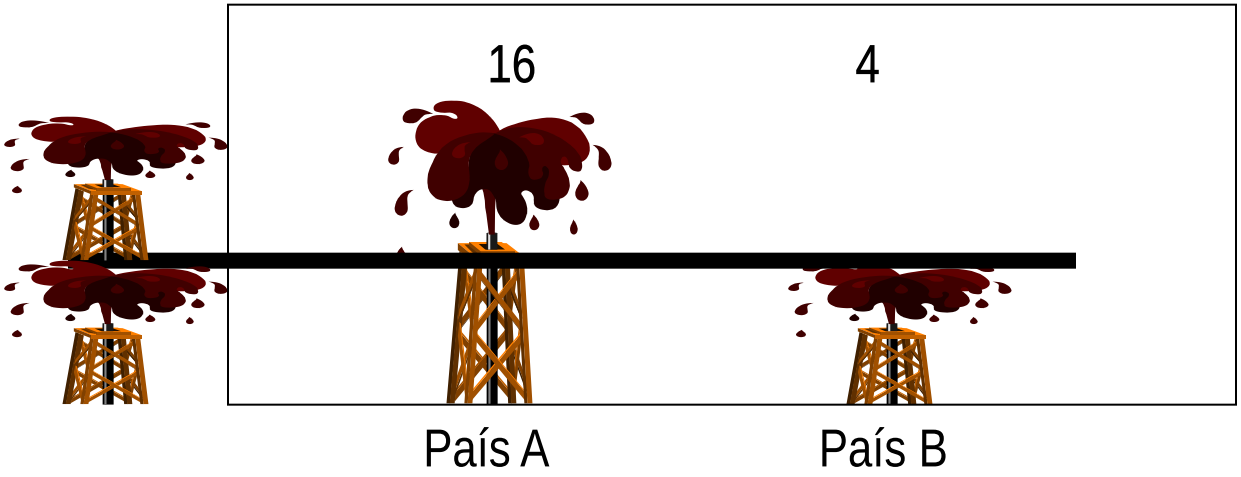
Distribución de frecuencias y representaciones gráficas

Diagrama/gráfico de barras:

Pictogramas: Representan, mediante un dibujo con alusión al tema de estudio, las frecuencias de las modalidades de una variable

El área debe ser proporcional a la frecuencia de la modalidad que representa, cuidado con diferencia entre el área y la altura

Gráfico III "Producción de petróleo en País A y B"

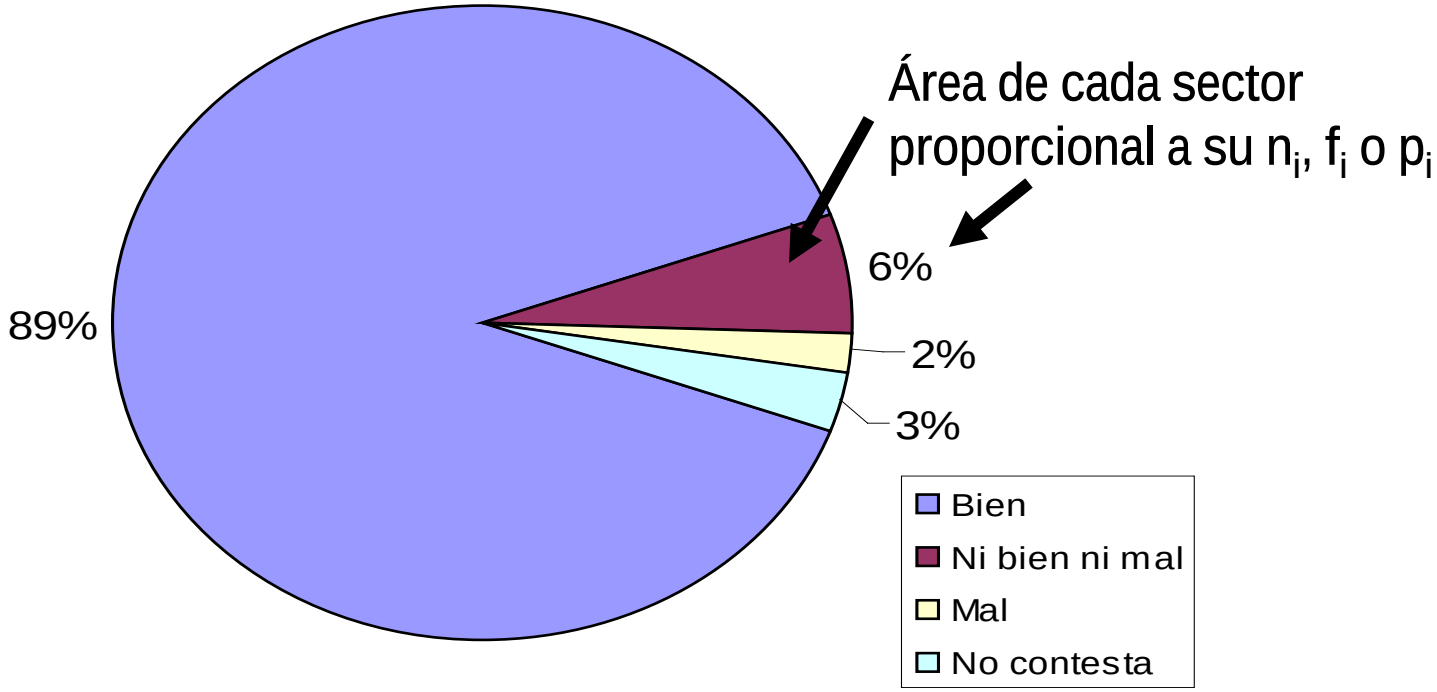


Fuente: elaboración propia

Distribución de frecuencias y representaciones gráficas

Diagrama de sectores

Gráfico II "A Ud. le parece bien o mal que en la misma clase hay alumnos y alumnas de distintos orígenes y culturas. Euskadi 2004



Fuente: Gobierno Vasco

Distribución de frecuencias y representaciones gráficas

Diagrama de sectores

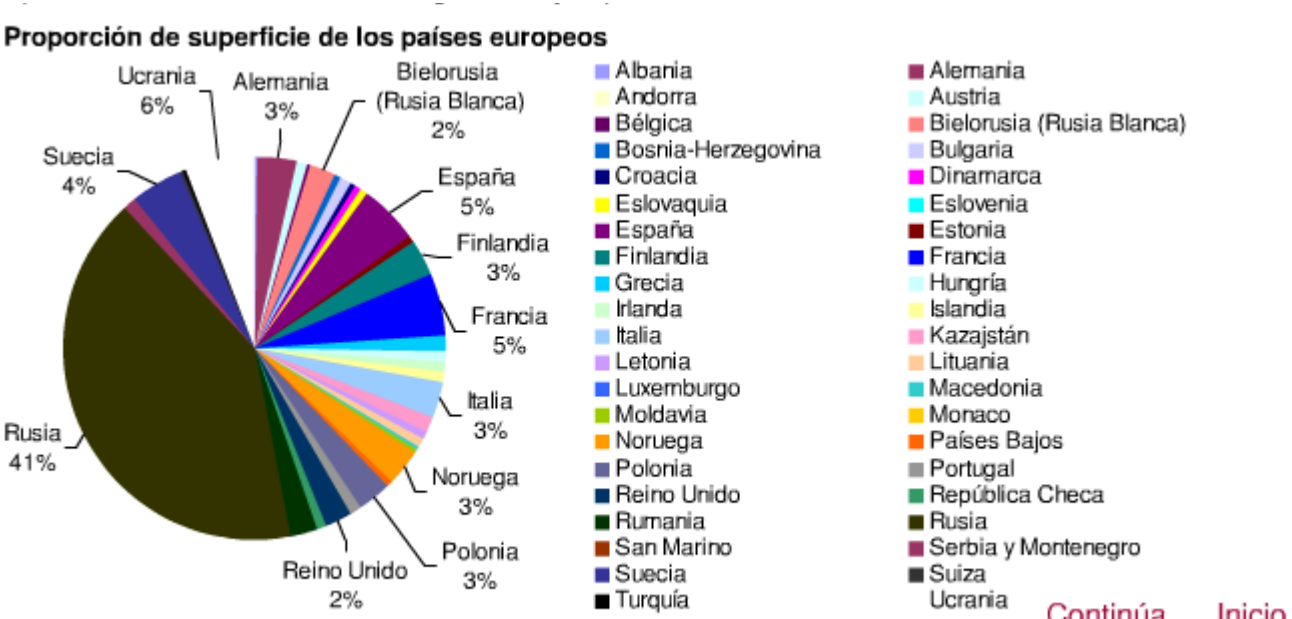
- Variables cualitativas o discretas

- Útil para comparación de las diferentes modalidades o para describir el peso de cada una de ellas en el total de la población

- No es útil cuando las modalidades de respuesta son muchas

Distribución de frecuencias y representaciones gráficas

Demasiadas modalidades de respuesta...



Ejemplo tomado de INE (*Explica: Tipo de gráficos ¿cuál uso?)*

Distribución de frecuencias y representaciones gráficas

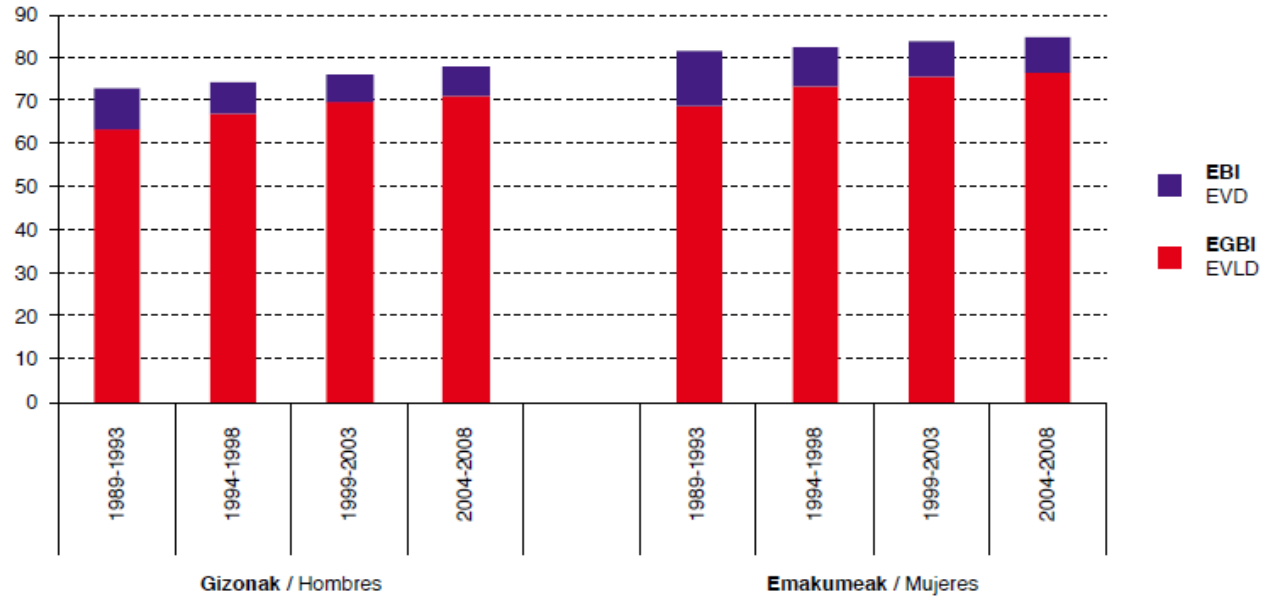
Tendencias temporales:

Se suelen representar mediante gráficos de barras o diagramas de líneas

Evolución de la esperanza de vida (total barra), esperanza de vida libre de discapacidad (EVLD) y esperanza de vida con discapacidad (EVD) al nacer por sexo en la C.A. de Euskadi

20

Bizi-itxaropenaren bilakaera (barra osoa), ezintasunik gabeko bizi-itxaropena (EGBI) eta ezinduta bizitzaren itxaropena (EBI) jaiotzean, sexuaren arabera, Euskal AEn



Iturria: Geuk egina, Eustaten oinarrituta.
Fuente: Elaboración propia a partir de Eustat.

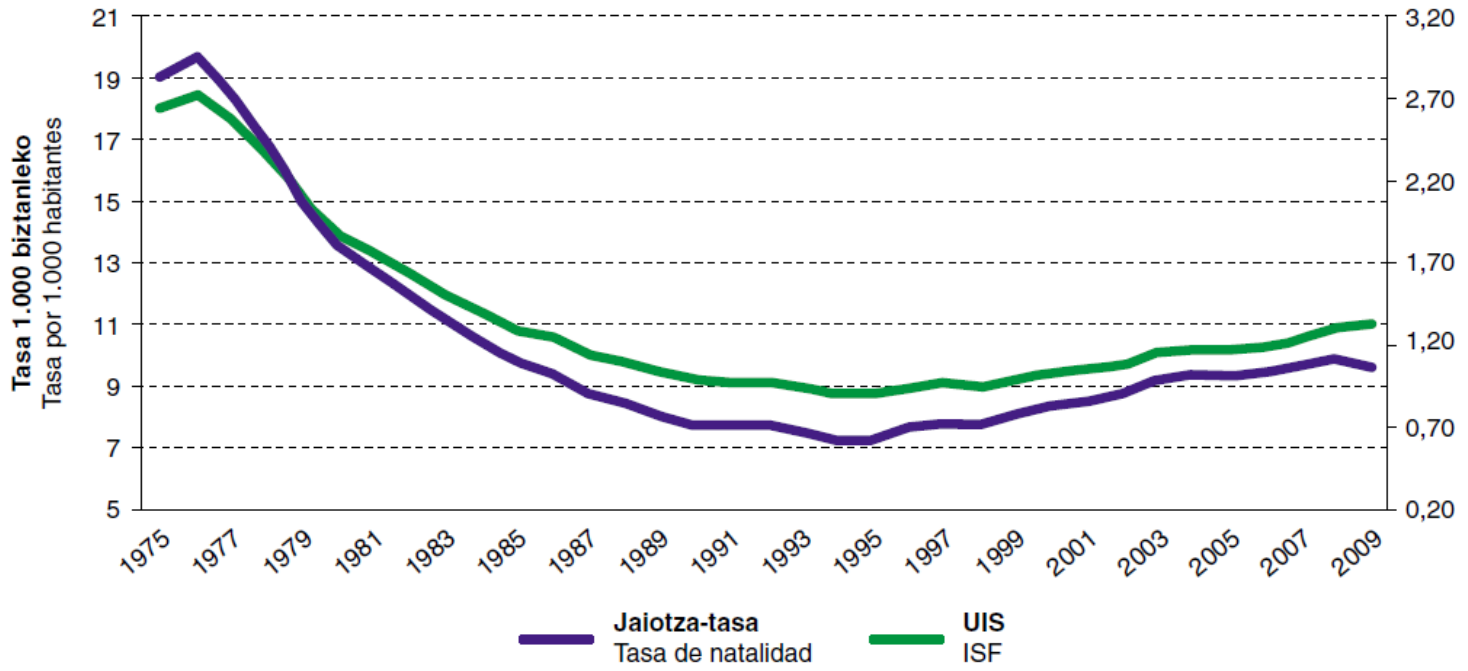
Distribución de frecuencias y representaciones gráficas

Tendencias temporales:

Evolución de la tasa de natalidad y el índice sintético de fecundidad (ISF) de la C.A. de Euskadi

5

Euskal AEren jaiotza-tasaren eta ugalkortasun-indize sintetikoaren bilakaera



Iturria: Geuk egina, Eustaten oinarrituta.
Fuente: Elaboración propia a partir de Eustat.

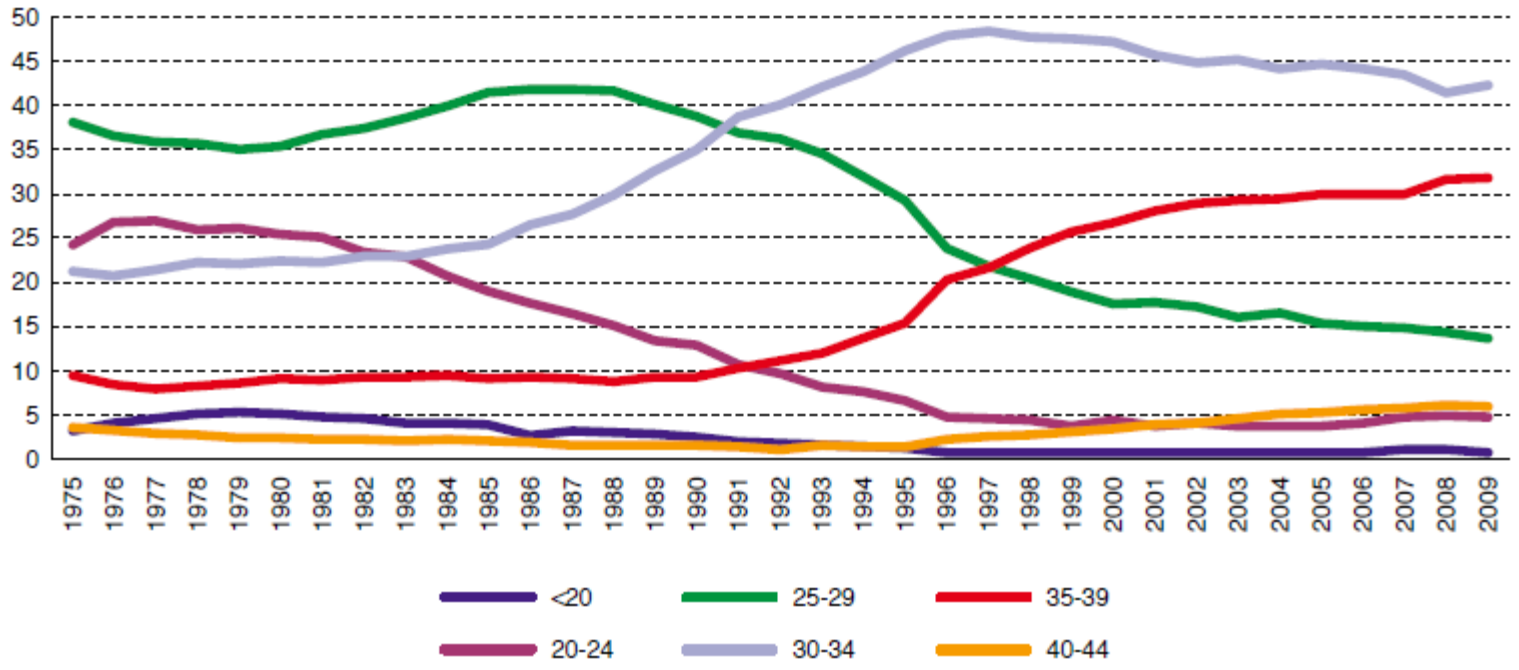
Distribución de frecuencias y representaciones gráficas

Tendencias temporales: comparación de tendencias

Evolución de la proporción de nacimientos por edad materna en la C.A. de Euskadi

7

Jaiotzen proportzioaren eboluzioa amaren adinaren arabera Euskal AEn

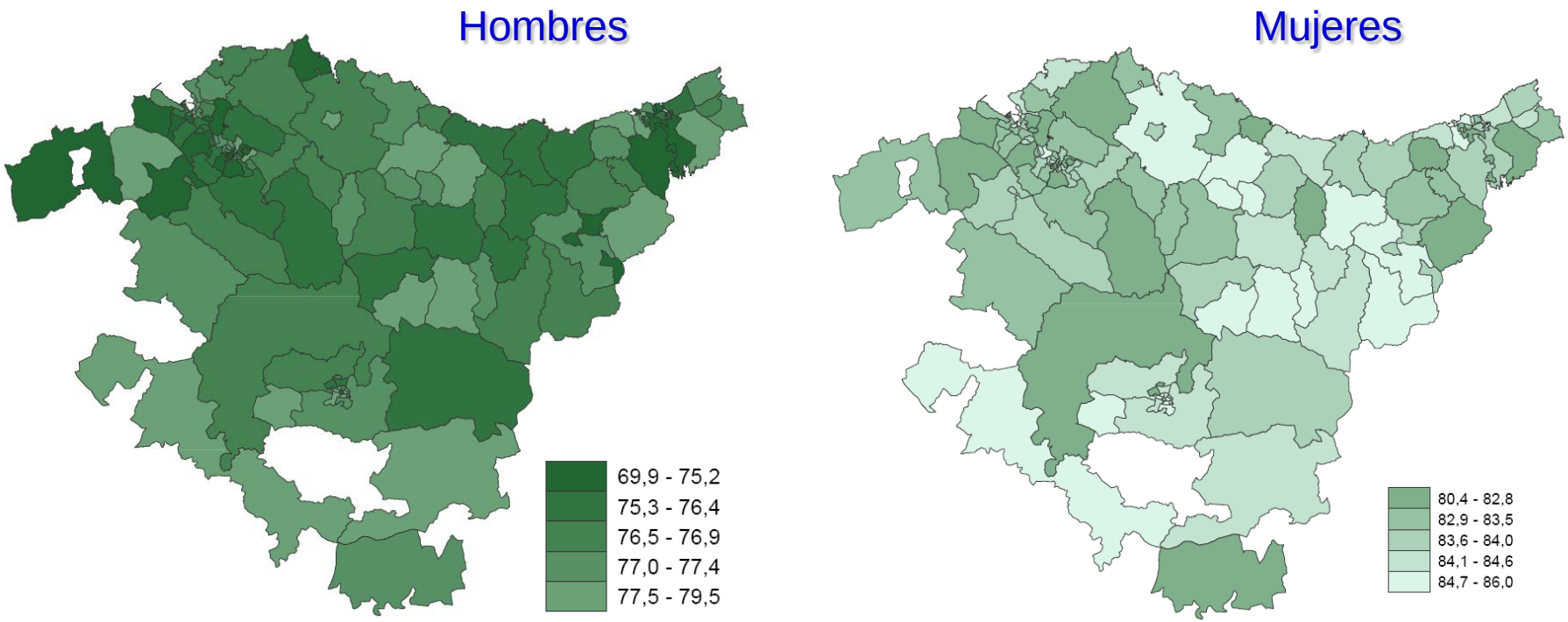


Iturria: Geuk egina, Eustaten oinarrituta.
Fuente: Elaboración propia a partir de Eustat.

Distribución de frecuencias y representaciones gráficas

Cartogramas

Esperanza de vida al nacimiento por zonas básicas de salud, 2001-2005



Fuente: Servicio de Estudios e Investigación Sanitaria. Departamento de Sanidad

Medidas descriptivas para variables cuantitativas

Una vez obtenido un resumen de la información mediante tablas y gráficos, nos interesa obtener medidas que resuman la información de la variable. Esta información **numérica** será básicamente:

- ✓Principales medidas de **tendencia central**: media, mediana y moda
- ✓Principales medidas de **posición**
- ✓Principales medidas de **dispersión (variabilidad)**: desviación típica, varianza y coeficiente de variación.
- ✓Principales medidas de **asimetría y forma**

Principales medidas de tendencia central: media, mediana y moda

a- **Media**, suma de todos los valores entre el total de individuos

- La más utilizada
- Centro de gravedad de la variable

Datos sin agrupar:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Ejemplo:

1,1,1,2,2,3,3,4,4,5

$$\overline{x} = \frac{1}{10}\left(1+1+1+2+2+3+3+4+4+5\right) = 2{,}6$$

Principales medidas de tendencia central: media, mediana y moda

Datos agrupados en una tabla:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i\, n_i = \sum_{i=1}^{k} x_i\, f_i$$

Ejemplo:

x_i	n_i	f_i
1	3	0,3
2	2	0,2
3	2	0,2
4	2	0,2
5	1	0,1

$$\bar{x} = \frac{1}{10}\left(1\cdot 3 + 2\cdot 2 + 3\cdot 2 + 4\cdot 2 + 5\cdot 1\right) = 2{,}6$$

$$\bar{x} = 1\cdot 0{,}3 + 2\cdot 0{,}2 + 3\cdot 0{,}2 + 4\cdot 0{,}2 + 5\cdot 0{,}1 = 2{,}6$$

Principales medidas de tendencia central: media, mediana y moda

Propiedades de la media:

1) $$y_i = x_i + k \Rightarrow \bar{y} = \bar{x} + k$$

1,2,3 media = 2

2,3,4 media = 3

2) $$y_i = x_i \cdot k \Rightarrow \bar{y} = \bar{x} \cdot k$$

1,2,3 media = 2

2,4,6 media = 4

Principales medidas de tendencia central: media, mediana y moda

3) $z_i = x_i + y_i \Rightarrow \bar{z} = \bar{x} + \bar{y}$

$$z_i = a \cdot x_i + b \cdot y_i \Rightarrow \bar{z} = a \cdot \bar{x} + b \cdot \bar{y}$$

4) $\sum_{i=1}^{N} (x_i - \bar{x}) = o$

1,2,3 media= 2

(-1)+0+1=0

La suma de las diferencias respecto a la media es igual a 0

Principales medidas de tendencia central: media, mediana y moda

5) Dados dos grupos A y B con tamaño N_a y N_b respectivamente, la media conjunta de una variable común es:

$$\overline{x} = \frac{Na \cdot \overline{x}a + Nb \cdot \overline{x}b}{Na + Nb}$$

Principales medidas de tendencia central: media, mediana y moda

El principal problema de la media es que está muy afectada por los valores extremos, por eso, no es adecuada en distribuciones no simétricas.

Ejemplo:

1,1,1,2,2,3,3,4,4,5	Media 2,6
1,1,1,2,2,3,3,4,4,527	Media 50,27

Principales medidas de tendencia central: media, mediana y moda

Otras medias: en función del tipo de variable u objetivos, hay otro tipo de medias que se pueden utilizar: geométricas, armónicas, cuadráticas

Media aritmética ponderada, todos los valores no tienen el mismo peso.

$$\overline{x}_p = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}$$

W_i= peso valor

Ejemplo:

Examen 7 puntos

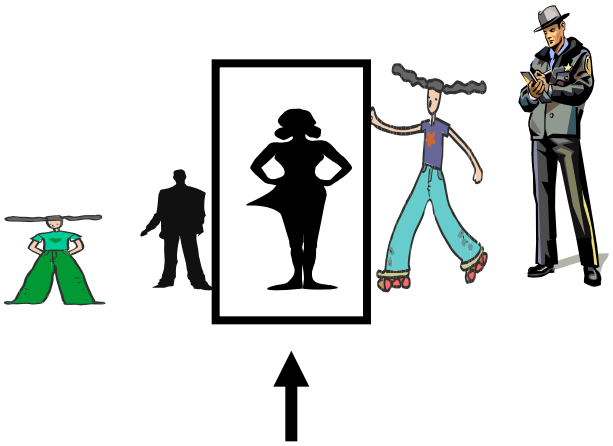
Trabajo 3 puntos

Con un 9 en examen y 8 en trabajo ¿Qué nota tengo?

$$\overline{x}_p = \frac{9 \cdot 7 + 8 \cdot 3}{7 + 3} = 8{,}7$$

Principales medidas de tendencia central: media, mediana y moda

b-Mediana: en una variable cuyas observaciones han sido ordenadas de menor a mayor, la mediana (M_{ed}) es el valor que deja por debajo y por encima el 50% de las observaciones.



El 50% de la muestra es igual o más alta/o

El 50% de la muestra es igual o más baja/o

Principales medidas de tendencia central: media, mediana y moda

¿Qué valor es la mediana?

Datos sin agrupar

Impares

1 [2] 3 El 50% ha sacado 2 o más en el test.

El 50% ha sacado 2 o menos en el test

Pares

1 [2 3] 4 $M_{ed} = \frac{2+3}{2} = 2{,}5$

El 50% ha sacado 2,5 o menos en el test

El 50% ha sacado 2,5 o más en el test.

Principales medidas de tendencia central: media, mediana y moda

Datos agrupados:

Impares

X_i	n_i	N_i
1	10	10
2	**20**	**30**
3	10	40
4	17	57
Total	57	

Con 57 observaciones el valor de la mediana será el que ocupe el orden 29 (57/2=28,5)

Orden	…	27	28	**29**	30	31
Valor	…	2	2	**2**	2	3

Principales medidas de tendencia central: media, mediana y moda

Datos agrupados:

Pares

x_i	n_i	N_i	F_i
1	10	10	0,2
2	**15**	**25**	**0,5**
3	**20**	**45**	**0,9**
4	5	50	1
Total	50		

Con 50 observaciones el valor de la mediana será el que ocupe el orden 25 y 26 (50/2=25)

Orden	…	24	25	**26**	27	28
Valor	…	2	2	**3**	3	3

M_{ed}= 2,5 (punto intermedio entre 2 y 3)

$$\frac{2+3}{2} = 2{,}5$$

Principales medidas de tendencia central: media, mediana y moda

Propiedades de la mediana

- No afectada por valores extremos (útil para asimétricas)

  2,5,7,9,11 $M_{ed}=7$

  2,5,7,9,527 $M_{ed}=7$

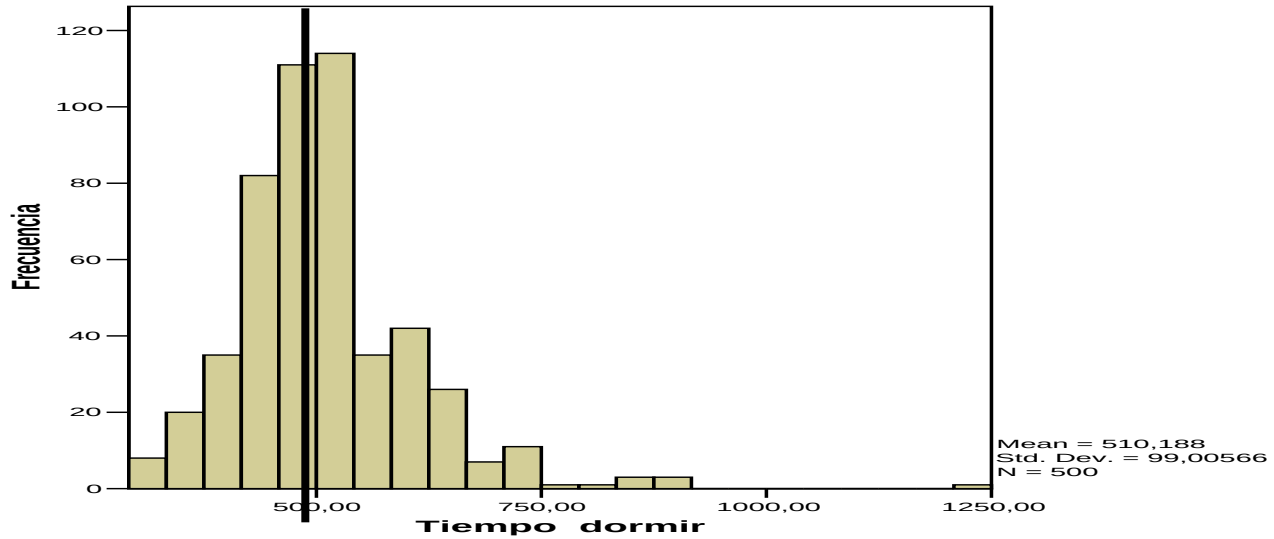
- Cálculo rápido e interpretación sencilla

- Siempre es un valor de la variable que estudiamos (importante en discretas)

Principales medidas de tendencia central: media, mediana y moda

- La mediana divide al histograma en dos partes iguales

"Tiempo dedicado a dormir los días laborables"



Principales medidas de tendencia central: media, mediana y moda

c-Moda: en una distribución de frecuencias, se denomina moda (M_0) al valor de la variable que más se repite (tiene mayor frecuencia)

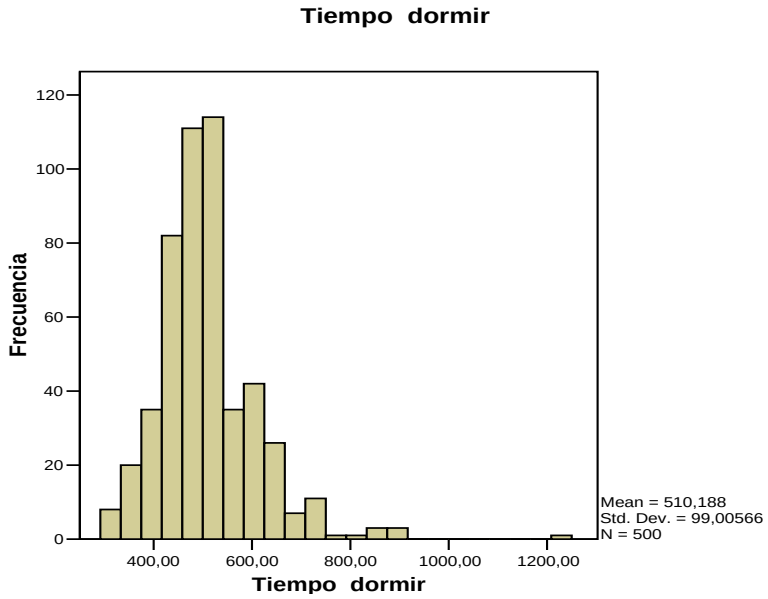
- Puede no haber moda
- Puede haber más de una moda (bimodal, trimodal…)
- También para variables cualitativas
- Útiles para distribuciones que se concentran en torno a un valor.

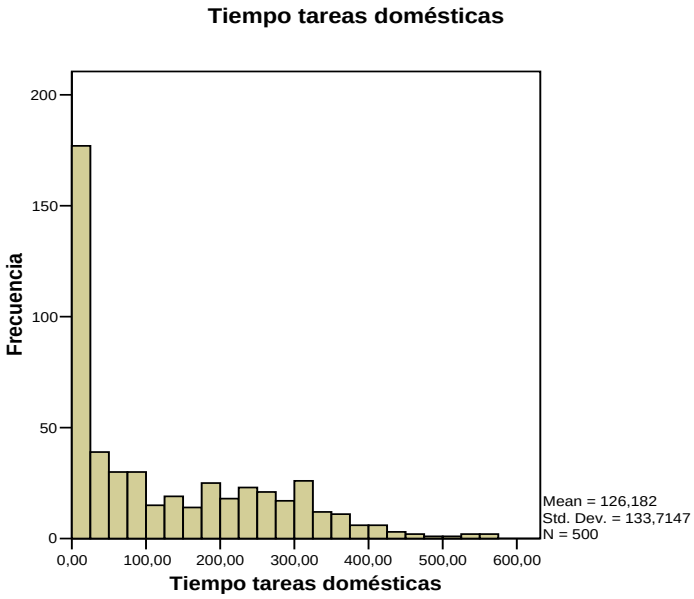
X_i	**n_i**
1	10
2	**20**
3	10
4	17
Total	57

X_i	n_i
Primarios	**317**
Secundarios	132
Superiores	51
Total	500

Principales medidas de tendencia central: media, mediana y moda

¿Qué estadísticos utilizarías para definir cada variable?





Estadísticos

	N					Percentiles		
	Válidos	Perdidos	Media	Mediana	Moda	25	50	75
Tiempo dormir	500	0	510,1880	495,0000	480,00	450,0000	495,0000	554,0000
Tiempo tareas domésticas	500	0	126,1820	75,0000	,00	,0000	75,0000	230,0000

Principales medidas de posición

En ocasiones nos interesa saber que posición ocupa un determinado dato/individuo en el total de la población.

El individuo 14 ha sacado 26 puntos en el test:

- ✓¿Es un buen resultado?
- ✓¿Qué posición ocupa ese individuo dentro de su clase?

Principales medidas de posición

Estas medidas indican la posición de un determinado individuo en la población, además sirven para dividir los datos en partes con igual frecuencia.

Cuartiles- Q_1, Q_2, Q_3 dividen en cuatro partes iguales

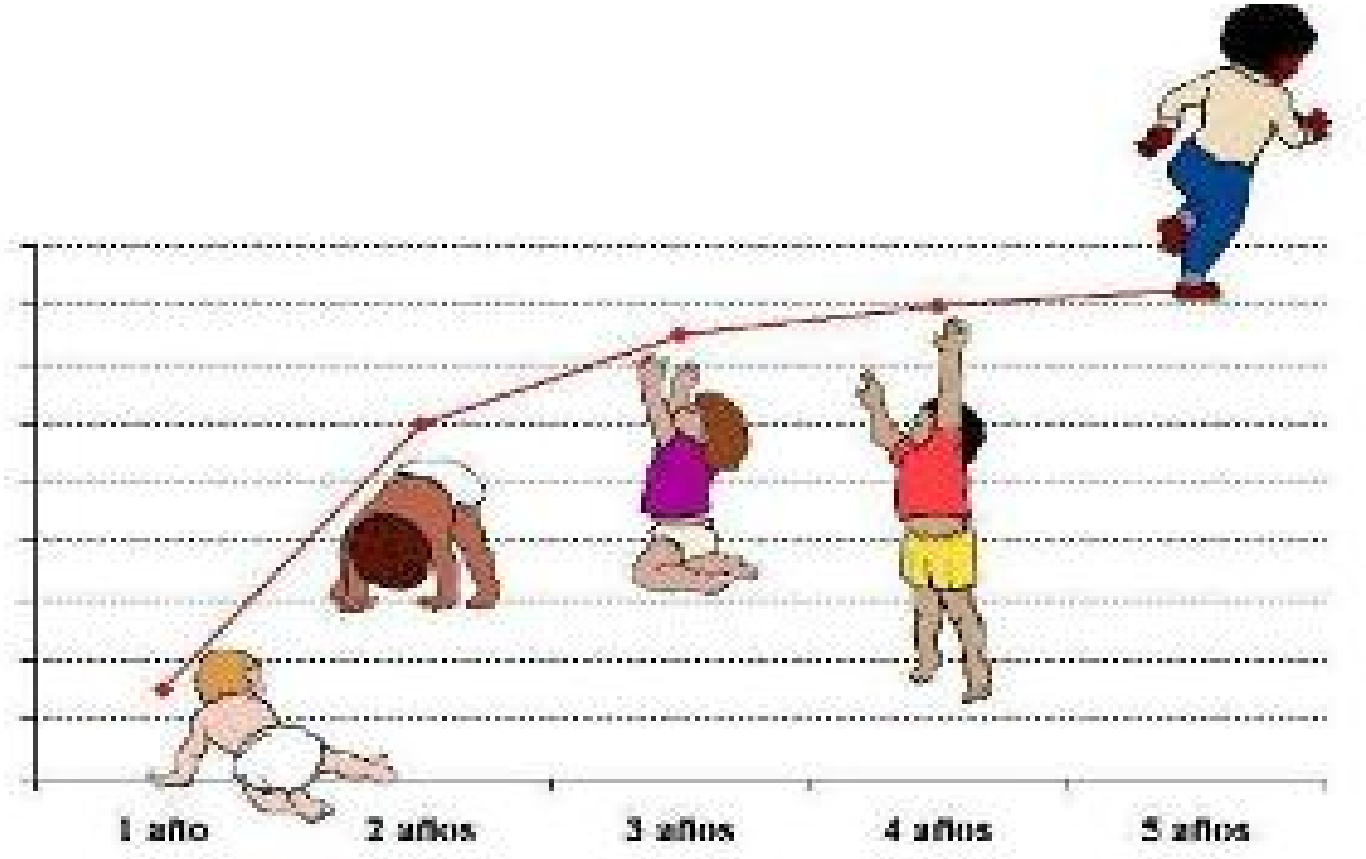
Deciles- $D_1,\ldots,D_9$ dividen en diez partes iguales

Percentiles- $P_1,\ldots.. P_{99}$ dividen en cien partes iguales

$$Q_2=D_5=P_{50}=M_{ed}$$

Principales medidas de posición

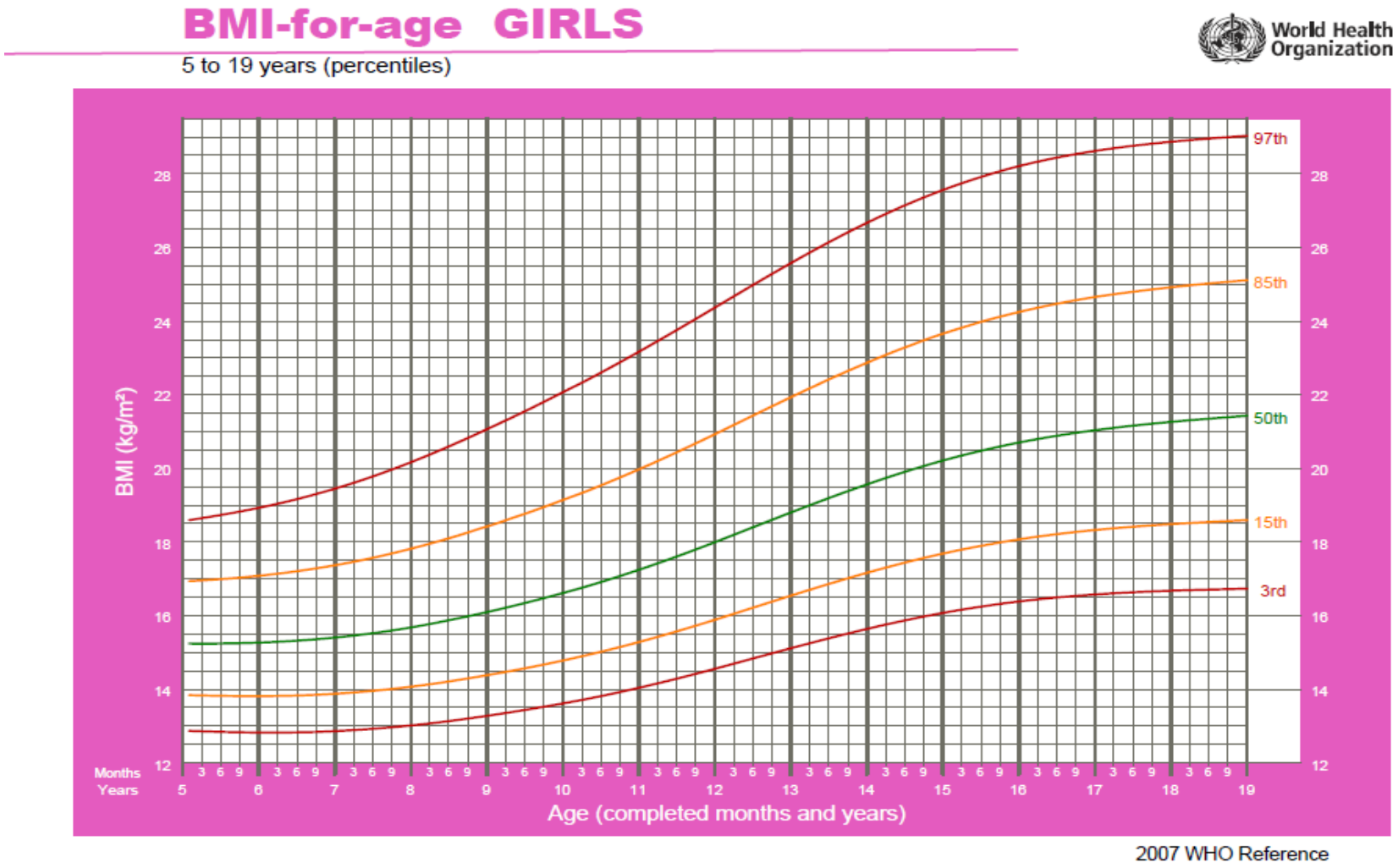
Los percentiles en la vida diaria…



¿Está mi niño gordito?/ Qué alta es mi niña! ¿O no?

Principales medidas de posición

Los percentiles en la vida diaría...



Principales medidas de posición

De los resultados obtenidos por una clase en el examen de estadística se sabe lo siguiente:

$Q_3 = 7$ $D_9 = 9$ $P_{40} = 5$

La puntación 7 deja por debajo al 75% de los/as alumnos/as y al otro 25% por encima

El 90% de la clase ha sacado un 9 o menos, y el 10% un 9 o más.

Al menos el 60% de la clase ha superado la asignatura

Sabemos que una determinada persona ha obtenido en un test una puntuación de 322. ¿Qué información nos aporta el hecho de que esa puntuación sea el P_{95} del total de la población?

Principales medidas de posición

Cálculo de percentiles

Variables discretas (datos agrupados y sin agrupar):

x_i	n_i	N_i	F_i
1	10	10	0,2
2	**15**	**25**	**0,5**
3	**20**	**45**	**0,9**
4	5	50	1
Total	50		

$P_{24} = 2$

24%= 12 ⟶ Posiciones 12º y 13º

¿Qué significa P24 = 2?

$D_3 = 2$

30% = 15, posiciones 15º y 16º

¿Qué porcentaje ha obtenido menos de un 2?

¿Qué porcentaje ha obtenido un 2 o más?

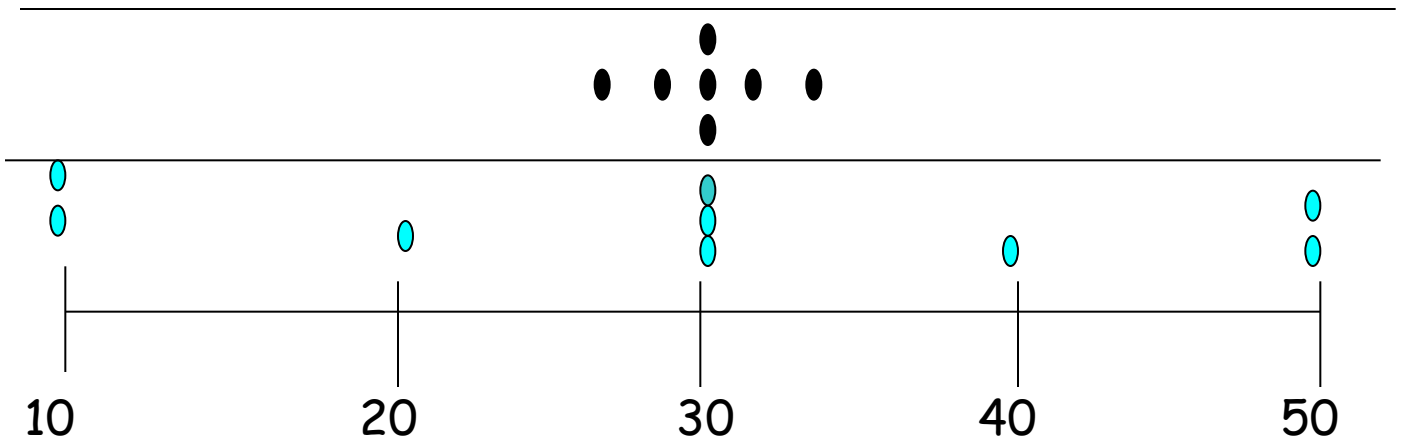
Medidas de dispersión

Las medidas anteriores nos son suficientes para describir la variable, tal y como demuestra lo siguiente:

A) 28/29/30/30/30/31/32 $\bar{x} = 30, M_{ed} = 30, M_o = 30$

B) 10/10/20/30/30/30/40/50/50 $\bar{x} = 30, M_{ed} = 30, M_o = 30$

A pesar de que las tres medidas son las mismas, los datos de B están mucho más dispersos. Por ello, son necesarios índices que nos den cuenta de esa dispersión



Medidas de dispersión

Las medidas de dispersión permiten analizar las diferencias que se dan entre los individuos en la variable, es decir, las diferencias entre los sujetos en las puntuaciones de la variable.

También permiten cuantificar la representatividad de una medida de posición, permitiéndonos establecer hasta qué punto una medida de tendencia central es representativa como síntesis de una distribución.

Medidas de dispersión:

Vamos a ver diferentes formas de describir la variabilidad de una variable:

1. Rango o recorrido
2. Recorrido intercuartílico y semi-intercuartílico
3. Diferencias respecto a la media:
    a) Desviación media
    b) Varianza
    c) Desviación típica
4. Coeficiente de variación
5. Diagrama de cajas

Medidas de dispersión

1. **Rango o recorrido**, es la variación total de los datos y se calcula restando a la puntuación mayor la menor (en ocasiones se le suma la unidad)

A) 32-28= 4

B) 50-10= 40

- Fácil de calcular y medida en unidades de la variable
- No utiliza todas las observaciones
- Muy afectada por los datos extremos

Medidas de dispersión

2. Recorrido intercuartílico y semi-intercuartílico

$$IQ = Q_3 - Q_1 \qquad Q = \frac{Q_3 - Q_1}{2}$$

- No tan afectado por valores extremos

Medidas de dispersión

3. Diferencias respecto a la media: desviación media, varianza y desviación típica.

Miden las distancias respecto a un punto central, en este caso la media, pero podría ser la mediana.

Desviación media:

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{j} |x_i - \bar{x}| \cdot n_i}{n}$$

Con el ejemplo anterior:

$$DM_A = \frac{|28-30| + |29-30| + \ldots + |32-30|}{7} = 0{,}857$$

$$DM_B = \frac{|10-30| + \ldots + |50-30|}{9} = 11{,}11$$

Medidas de dispersión

Varianza: media de las distancias al cuadrado.

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - (\bar{x})^2$$

$$S_A^2 = \frac{(28-30)^2 + \ldots + (32-30)^2}{7} = 1{,}43$$

$$S_B^2 = \frac{(10-30)^2 + \ldots + (50-30)^2}{9} = 200$$

Medidas de dispersión

Datos agrupados:

$$S_x^2 = \frac{\sum_{i=1}^{j} n_i \cdot (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{j} n_i \cdot x_i^2}{n} - (\bar{x})^2$$

x_i	n_i
1	2
2	2

$$\bar{x} = 1{,}5$$

$$S^2 = \frac{(1-1{,}5)^2 \cdot 2 + (2-1{,}5)^2 \cdot 2}{4} = 0{,}25$$

En ocasiones, al trabajar con muestra se utiliza la cuasivarianza, que es lo mismo pero dividido por n-1 y no por n

Medidas de dispersión

Desviación típica:

$$S_x = \sqrt{S_x^2}$$

- Mejor interpretación que la varianza

$$S_A = \sqrt{1{,}43} = 1{,}19$$

$$S_B = \sqrt{200} = 14{,}14$$

Medidas de dispersión

-Propiedades de la varianza y desviación típica:

-Son sensibles a la variación de algún dato.

-Tienen muchas propiedades para la estimación

-No son recomendables cuando la media no es una buena medida de tendencia central

-No pueden ser valores negativos

$$y_i = x_i + k \Rightarrow S_y^2 = S_x^2$$

$$y_i = x_i \cdot k \Rightarrow S_y^2 = S_x^2 \cdot k^2$$

Medidas de dispersión

Comparando la variabilidad de dos variables:

En una población se han obtenido las siguientes medidas respecto a la altura (centímetros) y peso (kilogramos) de sus individuos:

$$\overline{X}_p = 67; S_p = 20$$

$$\overline{X}_a = 176; S_a = 30$$

¿Hay más variabilidad en el peso o en la altura?

Medidas de dispersión

4. **Coeficiente de variación**: Resulta útil para comparar la variabilidad de dos variables medidas en diferentes escalas.

$$CV = \frac{S_x}{\overline{X}}$$

(En ocasiones se expresa en porcentajes)

$$CV_p = \frac{20}{67} = 0{,}298$$

$$CV_a = \frac{30}{176} = 0{,}17$$

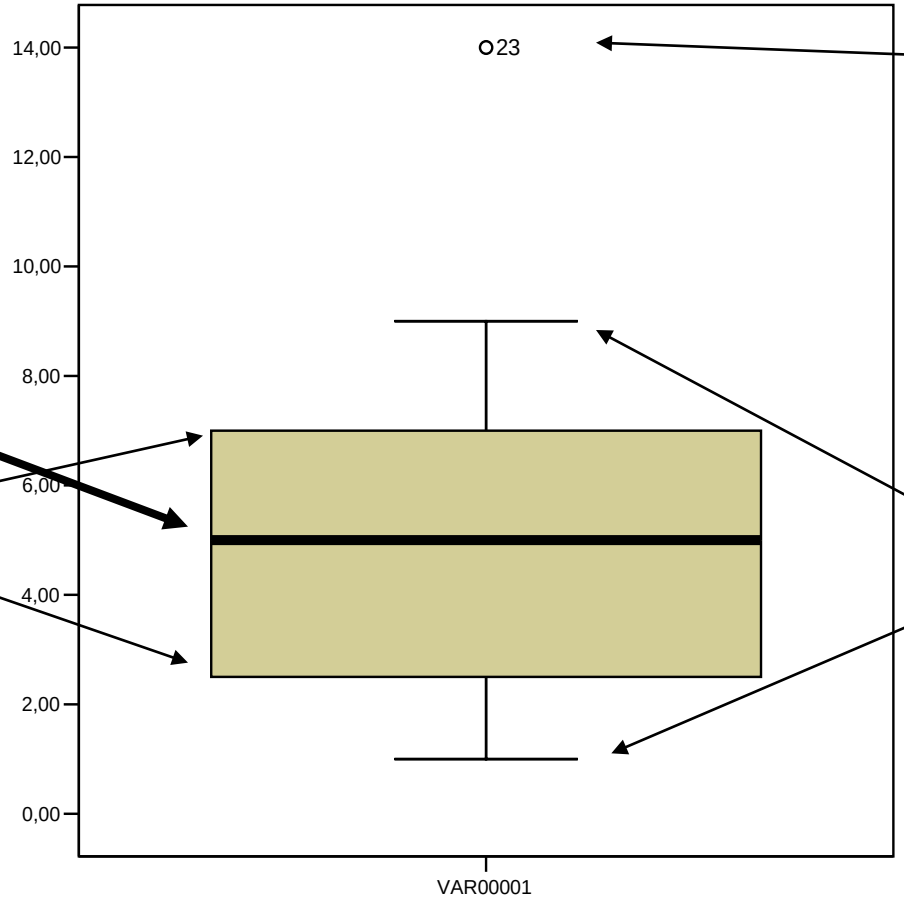
-Sólo para variables con datos positivos

-Invariable a los cambios de escala

Medidas de dispersión

5. Ayudas gráficas: el diagrama de cajas

Representa el valor central, mediana, y la variabilidad de la variable



Valor atípico

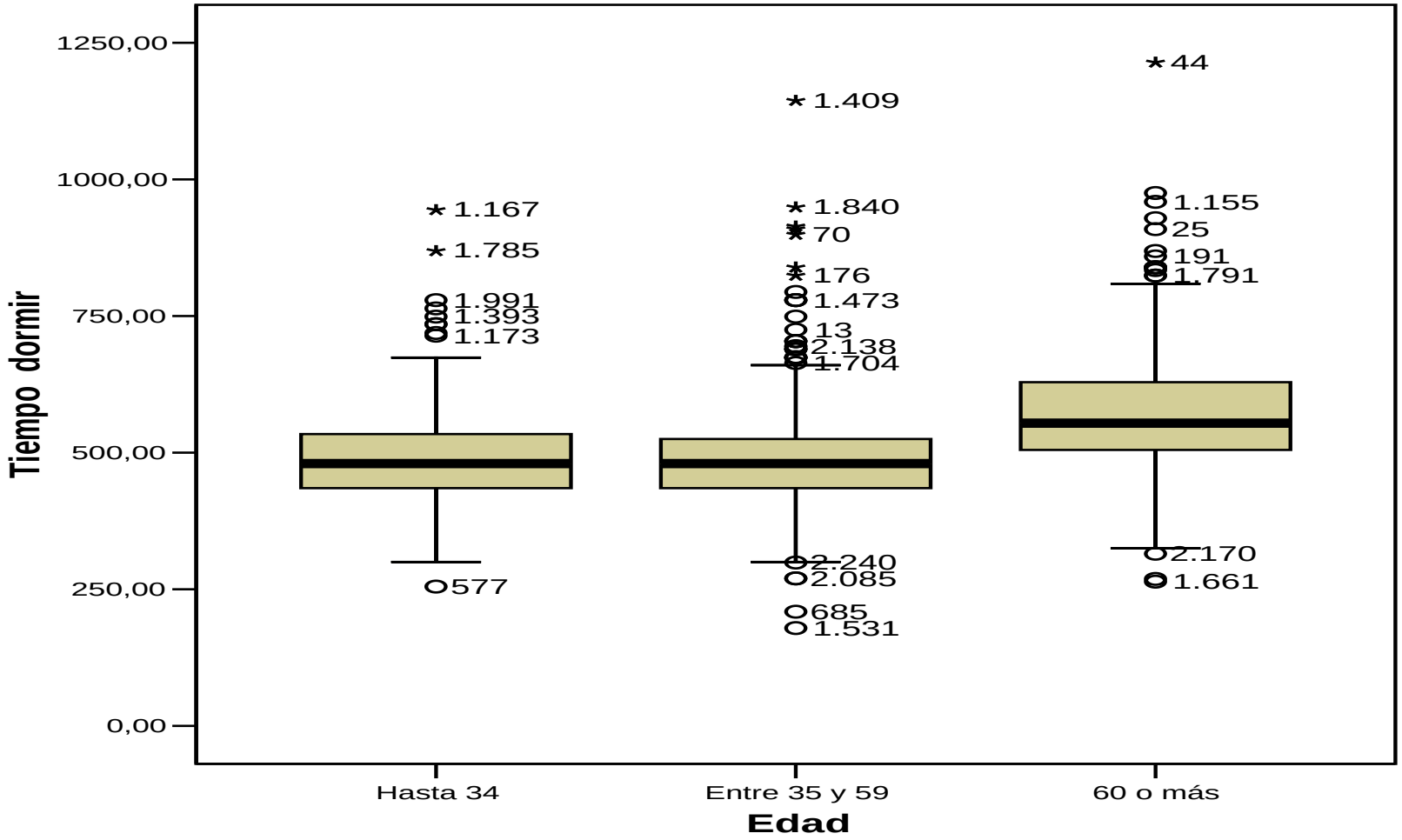
Se desvía 2S

Mediana

Q_1 y Q_3

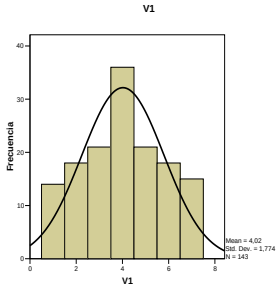
Valor mayor y menor

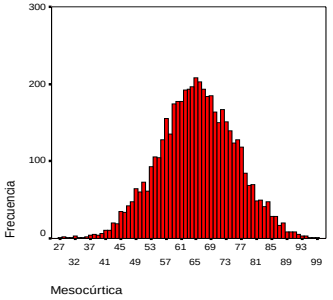
Medidas de dispersión



¿Cómo interpretarías estos diagramas de cajas?

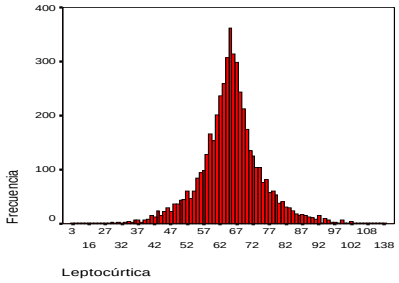
Principales medidas de forma y asimetría

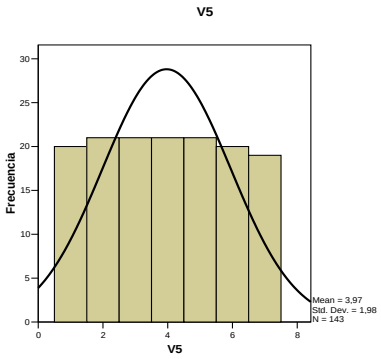




Nos ayudan a definir mejor la variable, nos muestran la forma de la distribución.

Nos permiten saber si los datos se reparten de una forma simétrica, y el nivel de apuntamiento de la distribución.

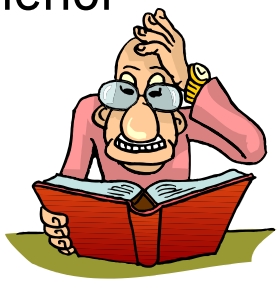




Principales medidas de forma y asimetría

Coeficiente de asimetría:

- Una distribución es simétrica si su parte derecha es igual a la izquierda (espejo)
- La asimetría es positiva o negativa dependiendo de donde este la cola de la distribución
- La media tiende a desplazarse hacia los valores extremos
  - ✓En distribuciones simétricas media, mediana y moda coinciden
  - ✓¿En asimétricas positivas, la media es mayor o menor que la mediana?



Principales medidas de forma y asimetría

Varias maneras de calcular el coeficiente de asimetría:

$$A_s = \frac{3\left(\overline{X} - M_e\right)}{S_x}$$

$$A_s = \frac{m_3}{S_x^3}; m_3 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3 \cdot n_i}{n}$$

A_s=0 Simétrica; A_s< 0 Asimétrica negativa; A_s>0 Asimétrica positiva

Principales medidas de forma y asimetría

valor cero → distribución simétrica
valor positivo → cola de la derecha más larga (asimetría + por la derecha)
valor negativo → cola de la izquierda más larga (asimetría - por la izquierda)

Principales medidas de forma y asimetría

Estadísticos

		V1	V2	V3
N	Válidos	143	143	143
	Perdidos	0	0	0
Media		4,02	5,20	-5,20
Mediana		4,00	4,00	-4,00
Moda		4	4	-4
Desv. típ.		1,774	3,492	3,492
Asimetría		-,001	1,161	-1,161
Error típ. de asimetría		,203	,203	,203
Curtosis		-,865	,676	,676
Error típ. de curtosis		,403	,403	,403

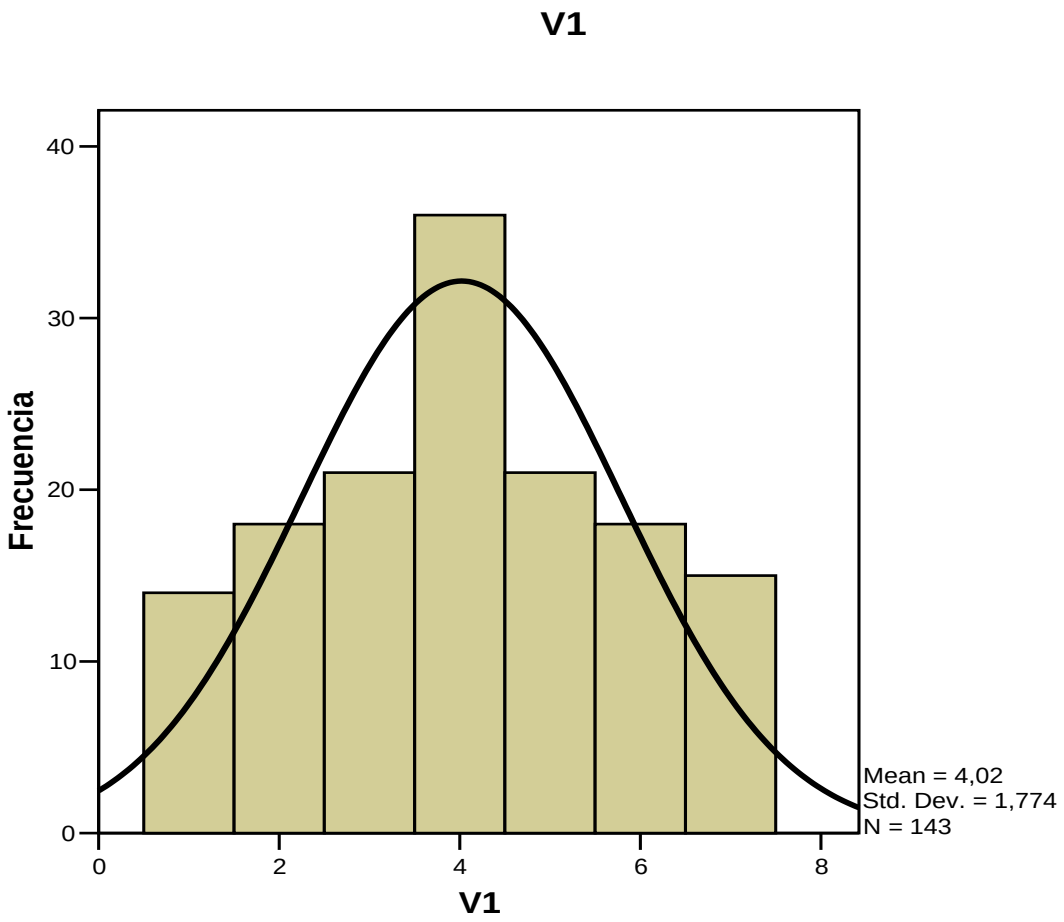
¿Cómo es la curva que dibuja el coeficiente de asimetría de cada variable?

Veamos….

Principales medidas de forma y asimetría

Estadísticos

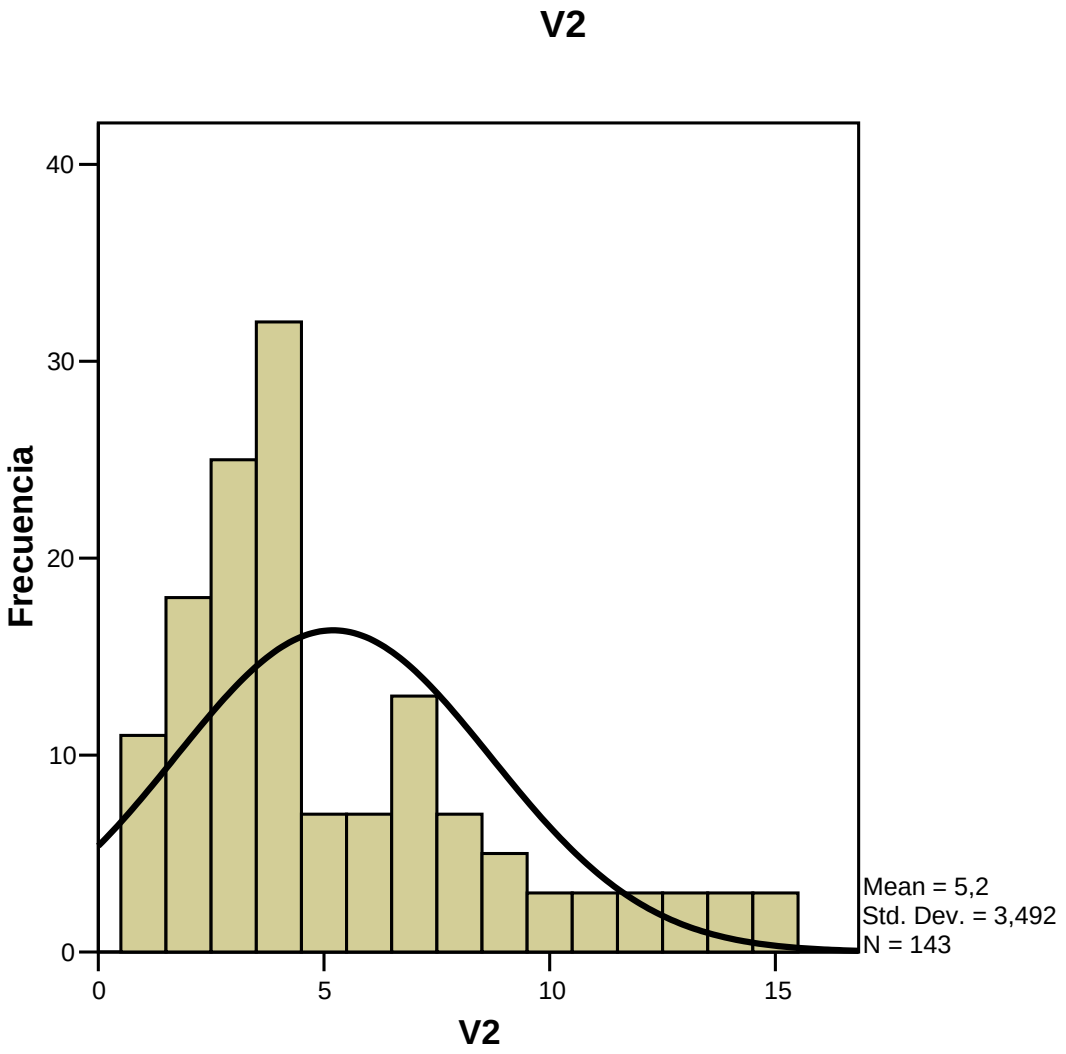
		V1	V2	V3
N	Válidos	143	143	143
	Perdidos	0	0	0
Media		4,02	5,20	-5,20
Mediana		4,00	4,00	-4,00
Moda		4	4	-4
Desv. típ.		1,774	3,492	3,492
Asimetría		-,001	1,161	-1,161
Error típ. de asimetría		,203	,203	,203
Curtosis		-,865	,676	,676
Error típ. de curtosis		,403	,403	,403



Principales medidas de forma y asimetría

Estadísticos

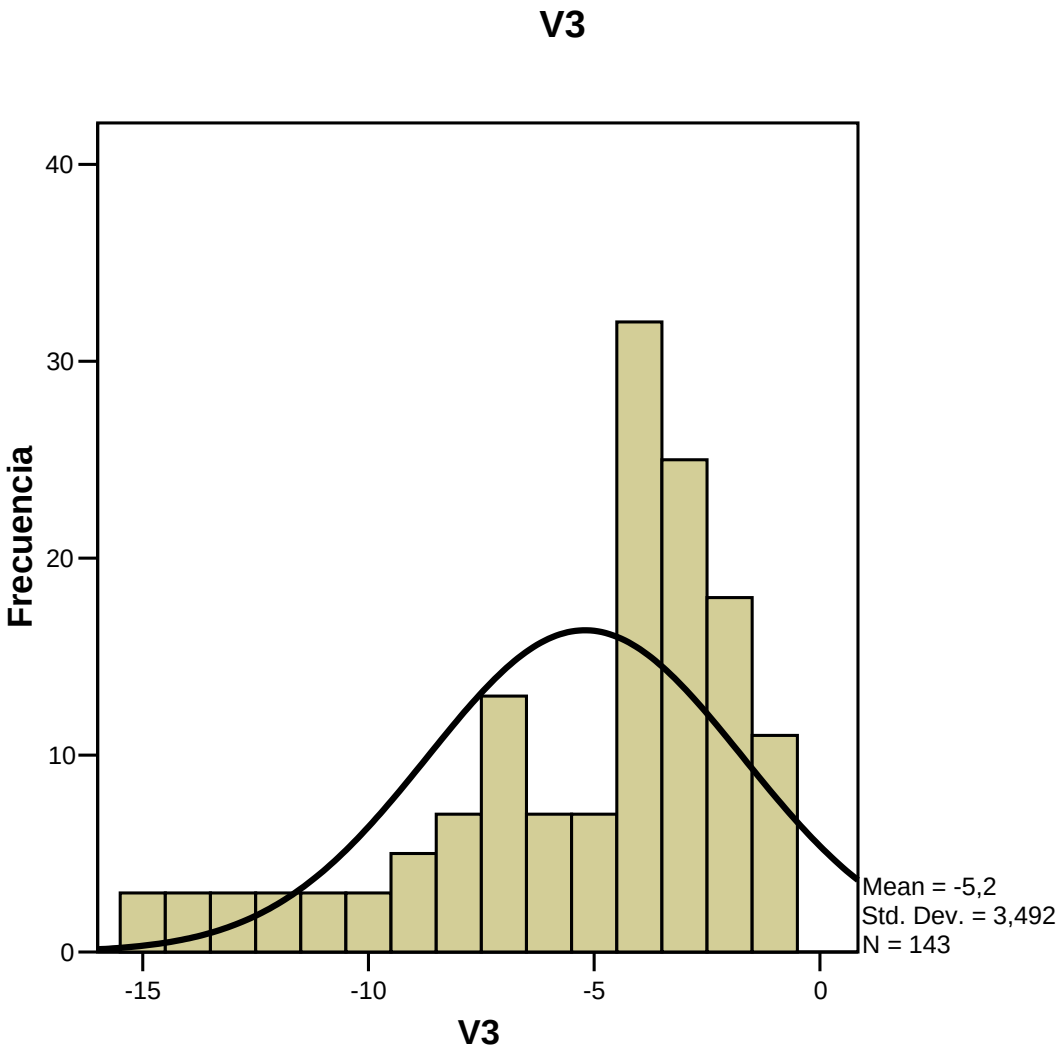
		V1	V2	V3
N	Válidos	143	143	143
	Perdidos	0	0	0
Media		4,02	5,20	-5,20
Mediana		4,00	4,00	-4,00
Moda		4	4	-4
Desv. típ.		1,774	3,492	3,492
Asimetría		-,001	1,161	-1,161
Error típ. de asimetría		,203	,203	,203
Curtosis		-,865	,676	,676
Error típ. de curtosis		,403	,403	,403



Principales medidas de forma y asimetría

Estadísticos

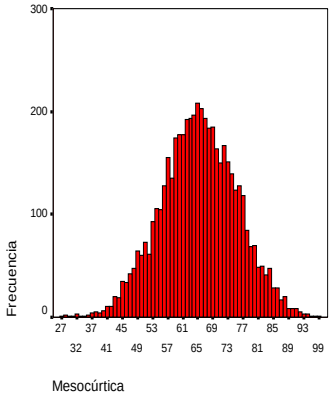
		V1	V2	V3
N	Válidos	143	143	143
	Perdidos	0	0	0
Media		4,02	5,20	-5,20
Mediana		4,00	4,00	-4,00
Moda		4	4	-4
Desv. típ.		1,774	3,492	3,492
Asimetría		-,001	1,161	-1,161
Error típ. de asimetría		,203	,203	,203
Curtosis		-,865	,676	,676
Error típ. de curtosis		,403	,403	,403



Principales medidas de forma y asimetría

Coeficiente de curtosis:

-Nos indica el grado de apuntamiento de una distribución comparada con una distribución normal. Mide si las observaciones se agrupan alrededor de un valor central o si existen valores alejados de la media



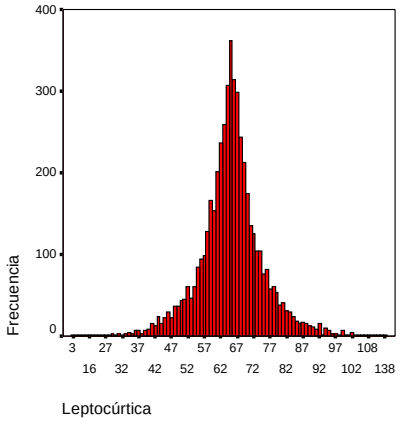
Mesocúrtica

$$a_4 = \frac{m_4}{S_x^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \cdot n_i}{\left(\frac{1}{n}\sum_{i=i}^{n}(x_i - \bar{x})^2 \cdot n_i\right)^2} - 3$$

Platicúrtica: curtosis < 0

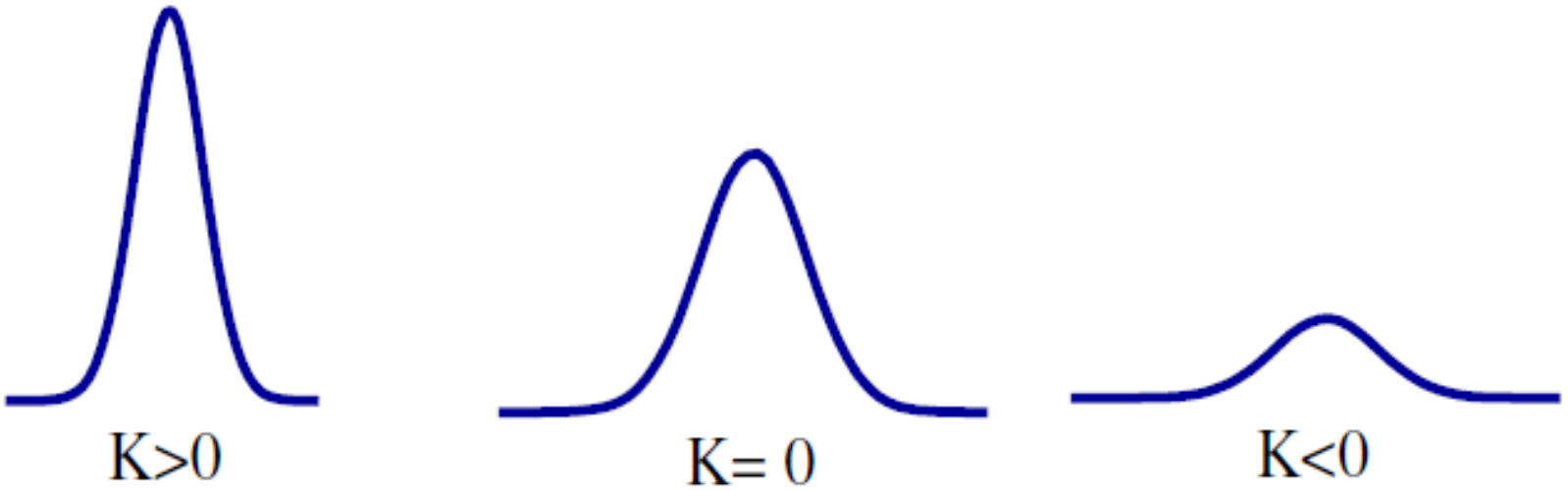
Mesocúrtica: curtosis = 0

Leptocúrtica: curtosis > 0



Leptocúrtica

Principales medidas de forma y asimetría



Principales medidas de forma y asimetría

Estadísticos

		V4	V5
N	Válidos	143	143
	Perdidos	0	0
Media		4,20	3,97
Mediana		4,00	4,00
Moda		4	2[a]
Curtosis		2,134	-1,222
Error típ. de curtosis		,403	,403

a. Existen varias modas. Se mostrará el menor de los valores.

¿Cómo es la curva que dibuja el coeficiente de curtosis de cada variable?

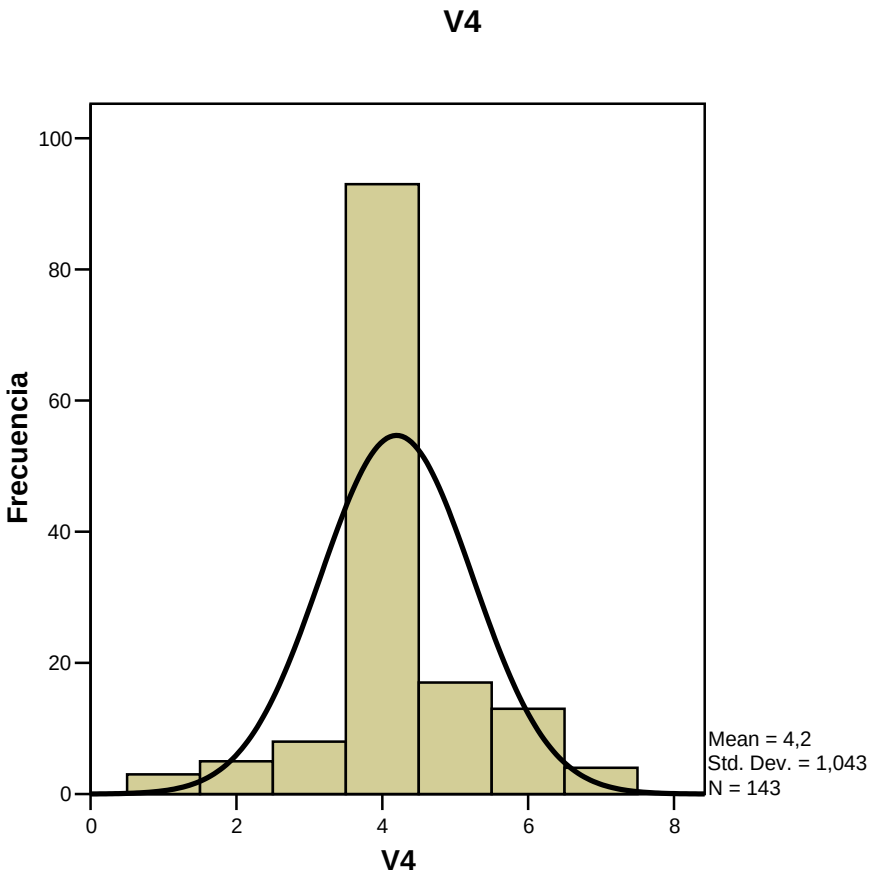
Veamos….

Principales medidas de forma y asimetría

Estadísticos

		V4	V5
N	Válidos	143	143
	Perdidos	0	0
Media		4,20	3,97
Mediana		4,00	4,00
Moda		4	2[a]
Curtosis		2,134	-1,222
Error típ. de curtosis		,403	,403

a. Existen varias modas. Se mostrará el menor de los valores.



Principales medidas de forma y asimetría

Estadísticos

		V4	V5
N	Válidos	143	143
	Perdidos	0	0
Media		4,20	3,97
Mediana		4,00	4,00
Moda		4	2[a]
Curtosis		2,134	-1,222
Error típ. de curtosis		,403	,403

a. Existen varias modas. Se mostrará el menor de los valores.