TERCERA PARTE:
CÓMO ANALIZAR Y DESCRIBIR LA VARIABILIDAD CUANDO LOS DATOS DIBUJAN UNA FORMA DE CAMPANA

Tema 5:

Análisis del sector central de la distribución: media y desviación estándar

Tema 5:

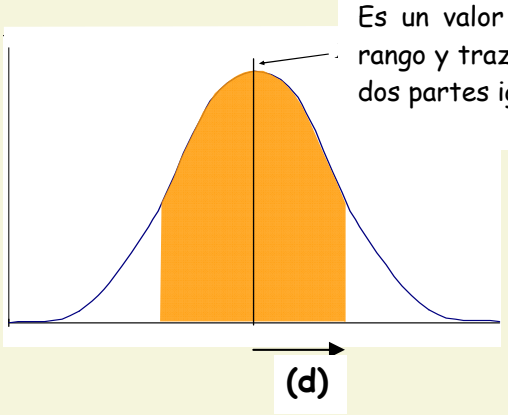
Análisis del sector central de la distribución: media y desviación estándar

INTRODUCCIÓN

Analizar el sector central en las distribuciones campaniformes significa determinar la variabilidad del conjunto de datos que aparecen agrupados en el sector central del rango. En los capítulos anteriores nos hemos referido a ellos como el grupo mayoritario de valores, para diferenciarlos del resto de valores que aparecen más dispersos hacia los extremos de la distribución.

Conceptualmente, el procedimiento para calcular la variabilidad del grupo mayoritario de valores es bien sencillo. Consta de dos pasos:

En un primer paso, calculamos el punto central de la distribución. Es un valor que, al situarlo en su lugar correspondiente en el rango y trazar una perpendicular a éste, divide la distribución en dos partes iguales.

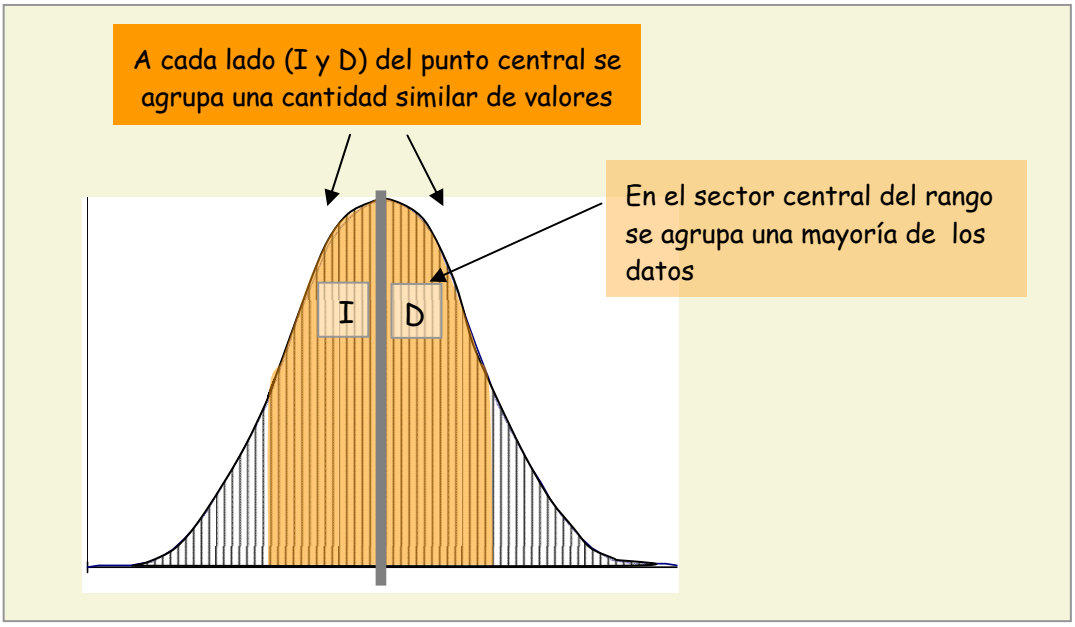


En un segundo paso, calculamos un valor que mide la distancia **(d)** desde el punto central al límite del intervalo que agrupa a la mayoría de valores de la variable.

EL PUNTO CENTRAL EN LAS DISTRIBUCIONES CAMPANIFORMES

El primer paso para hacer la descripción de la zona del rango que concentra la mayoría de los valores es calcular el punto central de la distribución, puesto que es el punto en torno al cual se agrupa la mayoría de los datos. Se trata de un valor que, en las distribuciones campaniformes, se sitúa en el sector central de la distribución.

En el gráfico adjunto, hemos dibujado una franja que se correponde con el sector central de la distribución. La línea en su interior, perpendicular a la línea del rango, representa el valor central al que nos referimos.



Existen diversos procedimientos para calcular el punto central de una distribución de valores. De todos los métodos existentes, el más ampliamente utilizado es, sin duda, la media aritmética o promedio. La facilidad de su cálculo explicaría el uso tan extendido de este procedimiento, incluso cuando dicho uso no es aconsejable.

Aunque su cálculo es una tarea muy sencilla que, además, hoy en día realizan los ordenadores, lo que no parece tan sencillo es comprender que no siempre es aconsejable usar el promedio para calcular el punto central de una distribución.

Como punto de partida, y como principio general también, podemos decir que el promedio sólo debe utilizarse para la descripción de variables cuya distribución sea campaniforme. En las páginas que siguen se darán las explicaciones que permiten entender esta afirmación.

PROMEDIO O MEDIA ARITMÉTICA

El promedio o media aritmética es el valor que se obtiene al sumar todos los datos y dividir el resultado entre el número total de datos.

Cálculo del promedio: fundamentos de la notación estadística

En todas las explicaciones que hemos presentado hasta ahora no hemos utilizado ninguno de los símbolos estadísticos habituales en cualquier manual. De aquí en adelante, será inevitable recurrir a dichos símbolos a la hora de presentar determinados procedimientos y fórmulas. Por este motivo, y porque entender y utilizar los símbolos estadísticos es algo mucho más fácil de lo que puede parecer en un principio, utilizaremos el espacio dedicado al promedio para explicar con detenimiento el significado de los símbolos estadísticos más comunes.

Para referirnos de forma genérica a cualquier variable que analizamos mediante las herramientas estadísticas, utilizamos el símbolo de la X, en mayúscula. **Lo que se hace, realmente, es utilizar una letra para mencionar de forma abreviada un concepto**. Supongamos, a modo de ejemplo, la siguiente expresión:

Se deben sumar todos los valores de la variable

Ahora, sustituiremos la palabra *variable* por su correspondiente símbolo:

Se deben sumar todos los valores de X

Expresado de este modo, se entiende que, sea cual sea la variable de la que estemos hablando, se deben sumar todos sus valores

Para mencionar los valores de la variable de cualquier población estadística utilizamos también como símbolo la x, esta vez minúscula, seguido de un subíndice que indica que se trata de cualquier valor de la variable: x_i

Si tomamos cualquier valor de la variable…

Si tomamos x_i

$$x_i$$

Cuando queremos concretar la mención y referirnos a un grupo de valores de cualquier variable se sustituye el subíndice (i) por números. Supongamos la expresión:

Tomaremos los seis primeros valores de la variable y…

Tomaremos $x_1\ x_2\ x_3\ x_4\ x_5\ x_6$ *y…*

$$x_1\ x_2\ x_3\ x_4\ x_5\ x_6$$

Siguiendo esta lógica, cuando queremos decir que una operación estadística consiste en sumar los valores de la variable, podemos escribirlo así:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \ldots\ldots$$

Supongamos ahora que debemos expresar mediante símbolos que hay que sumar todos los valores de la variable. Hacerlo como hasta ahora sería imposible porque cuando nos estamos refiriendo a cualquier variable, de forma genérica, el número de valores puede ser también cualquiera. Necesitamos, por tanto, una expresión que indique, de forma abreviada, **el concepto de *todos los valores de la variable*.**

Un modo de hacerlo es escribir, como anteriormente, unos cuantos valores de la variable con el subíndice correspondiente. Después se añaden puntos suspensivos y una última x, pero esta vez con el subíndice n, para indicar que sumaremos todos los valores de la variable que estamos analizando:

$$X_1, + X_2 + X_4 + X_5 \ldots\ldots X_n$$

Es la expresión …….Xn, la que indica que la suma continuará hasta llegar al último valor de la variable. La letra **n** alude siempre al número de valores que estamos analizando. Nos permite mencionar mediante un símbolo el

concepto de *número total de valores*, sin importar cuál sea esta cantidad.

Una vez explicado el significado de los símbolos más habituales, podemos presentar la expresión matemática de la media aritmética o promedio, que sustituye a su definición:

$$\overline{X} = \frac{X_1, + X_2 + X_4 + X_5 ..... X_n}{n}$$

La expresión matemática, que ahora debe resultarnos comprensible, se puede abreviar más todavía, ya que existen otros símbolos que lo permiten.

Tenemos, en primer lugar el símbolo Σ denominado ***sumatorio***. Colocado delante de una serie de números, indica que se deben sumar todos los números que hay a su derecha. Lo que nos permite es sustituir todos los símbolos de suma (+) por uno único.

Podemos simplificar la operación, y sustituir todos los símbolos de la suma por un único símbolo, y las menciones individuales de los valores por una única mención:

$$X_1, + X_2 + X_4 + X_5 .... X_n \longrightarrow \sum x_i$$

A la expresión que hemos escrito le faltaría indicar que la suma debe incluir todos los valores de la variable.

$$\sum_{i=1}^{n} x_i$$

$\sum x_i$

El símbolo del sumatorio seguido del símbolo de variable, indican que se sumarán los valores de la variable

$\sum_{i=1}^{n}$

Debajo del sumatorio se detalla el significado del subíndice i que acompaña a la x. Se indica, concretamente, que el subíndice alude a todos los valores de la variable, desde el primero (i=1) al último (i=n)

Finalmente, la fórmula matemática de la media queda así:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

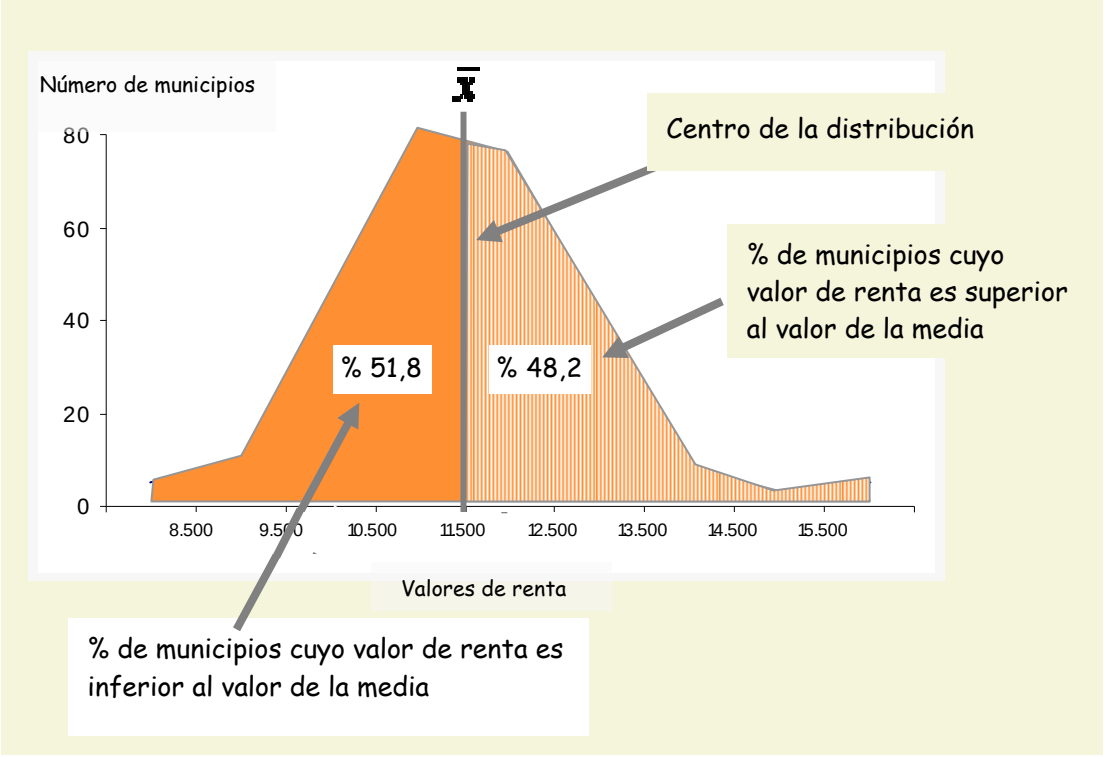
Utilización de la media

La media de los valores de una variable se utiliza para definir el punto central de distribuciones campaniformes. Como se ha indicado anteriormente, cuando la distribución tiene forma de campana una mayoría de los valores se agrupa a cada uno de los lados del punto central de la distribución.

Para entender mejor el significado y la utilización de la media veremos, a continuación dos ejemplos: uno de ellos se corresponde con una distribución campaniforme y, el otro, con una distribución de forma distinta, no campaniforme.

Veremos en primer lugar el ejemplo correspondiente a la distribución campaniforme. Los datos corresponden a los valores de renta personal disponible de los municipios de Euskadi en 2003.

La gráfica nos muestra lo que decíamos al inicio del tema:



En las distribuciones campaniformes una mayoría de los valores se concentra en el sector central del rango; en el centro de este sector, también, se sitúa el valor de la media

En relación a la variable de renta disponible, las características que comentamos se concretan en el hecho de que el valor de la media, 11.506 €, se sitúa en el

centro de la distribución; el porcentaje de municipios con valores inferiores y superiores a la media es similar.

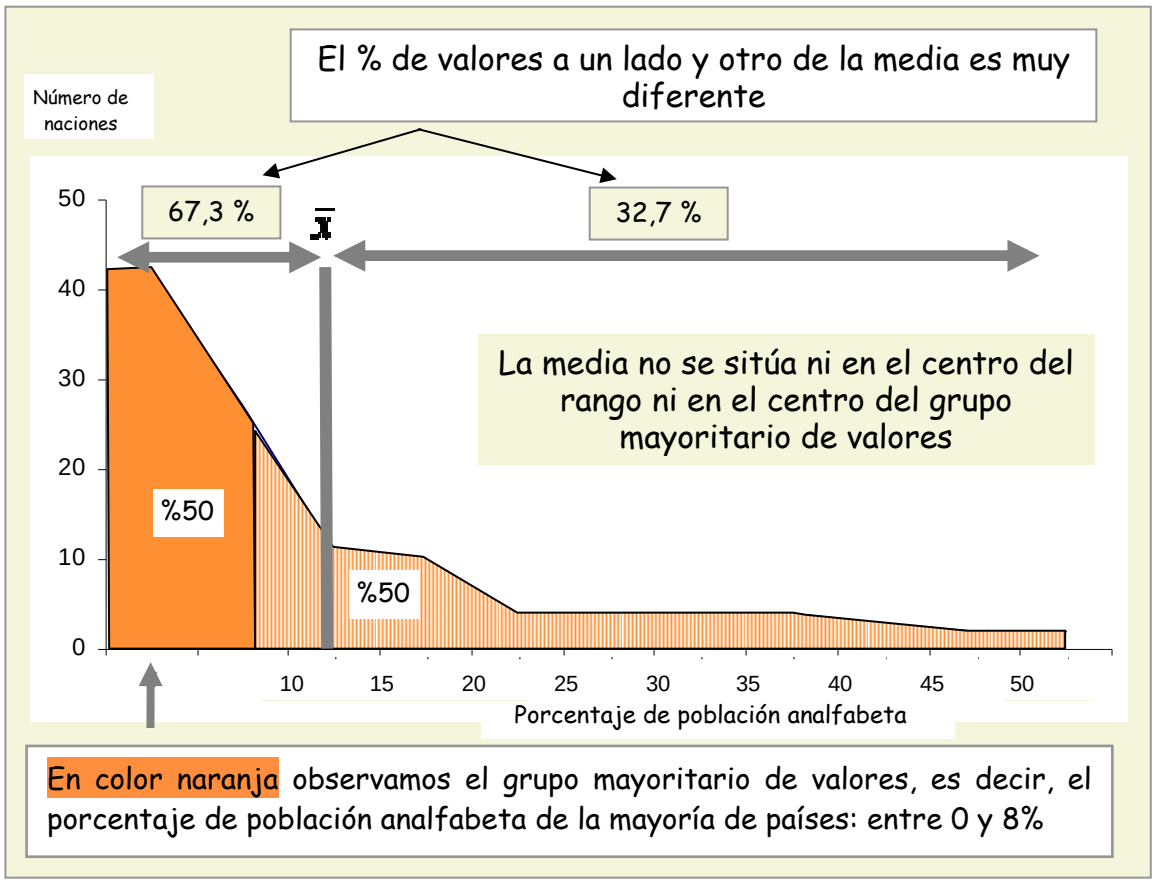
La conclusión fundamental que se extrae de lo que acabamos de mostrar es que:

> El valor de la media en las distribuciones campaniformes es un dato significativo y adecuado como descriptor de la distribución.

El valor de renta media de los municipios de Euskadi en 2003, 11.506 €, es un dato que representa de forma idónea al conjunto de valores de renta analizados: un número importante de municipios tuvieron en 2003 un valor de renta próximo al de la media. En la tabla anexa podemos observar que 193 municipios tuvieron un valor de renta comprendido entre los 10.000 y los 13.000 €. Estas 193 entidades constituyen el 77% del total de los municipios de la Comunidad Autónoma.

Intervalos	Nº de municipios	Porcentaje	Porcentaje acumulado
>10.000 ≤10.500	29	12	12
> 10.500 ≤ 11.000	27	11	23
> 11.000 ≤ 11.500	47	19	41
> 11.500 ≤ 12.000	37	15	56
> 12.000 ≤ 12.500	32	13	69
> 12.500 ≤ 13.000	21	8	77
	193	77 %	

A continuación veremos el ejemplo de una distribución no campaniforme. La población en este caso está constituida por 113 naciones del mundo, con niveles de desarrollo muy distintos. La variable que se analiza es el porcentaje de población analfabeta existente en cada uno de dichos países, según datos de 2006 recogidos por la UNESCO. (Ver Anexo nº 8) Como en el caso anterior, la gráfica nos muestra la distribución de valores dentro del rango. Mediante colores distintos hemos señalado también los dos lados de la distribución, cada uno de los cuales agrupa el 50% de los valores.



Resulta evidente que ahora no estamos ante una distribución campaniforme. La cuestión es ahora definir cuáles son las diferencias, además de la forma, entre las características de las dos distribuciones. Dicho de otro modo, se trata de relacionar las distintas formas de las distribuciones con las características de las poblaciones a las que representan.

- En la distribución no campaniforme, el grupo mayoritario de valores no se sitúa en el centro de la distribución sino en el extremo inferior de ésta. Esto significa, como podemos ver en la gráfica, que la mitad de los países tienen un porcentaje de población analfabeta inferior al 8%.

- El valor de la media, 12,2%, no se sitúa en el centro del grupo mayoritario de valores sino desplazado hacia la derecha en relación a aquellos. El hecho es que, mientras según el valor del promedio el porcentaje medio de población analfabeta se sitúa en un 12,2%, la gráfica muestra claramente que en una mayoría de países este porcentaje es claramente inferior. Del total de

naciones analizadas (113), un 67,3% tenían en 2006 un porcentaje de población analfabeta inferior al que indica la media.

La conclusión más importante que podemos extraer del ejemplo sobre la distribución del porcentaje de población analfabeta en los distintos países es que:

En las distribuciones no campaniformes, la media no es un valor representativo de la población, no proporciona una imagen idónea del centro de la distribución.

Intervalos	Nº de naciones	Porcentaje	Porcentaje acumulado
>0 ≤ 5	42	37,17	37,17
> 5 ≤ 10	27	23,89	61,06
> 10 ≤ 15	11	9,73	70,80
> 15 ≤ 20	10	8,85	79,65
> 20 ≤ 25	4	3,54	83,19
> 25 ≤ 30	4	3,54	86,73
	98		

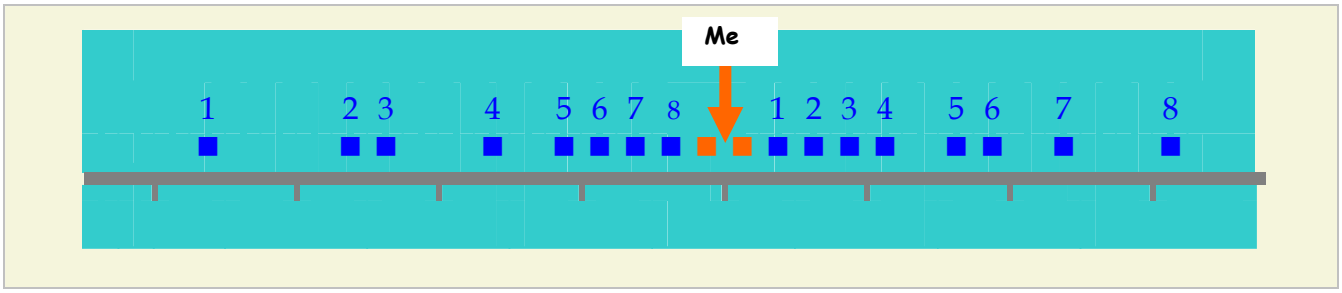
En un 61,06 % de los países, la población analfabeta no alcanza el 10%

¿Qué ocurriría entonces si utilizáramos el promedio como descriptor del porcentaje de población analfabeta en el mundo?. La respuesta es que proporcionaríamos una imagen distorsionada de la realidad que llevaría a pensar que, en la mayoría de los países, el porcentaje de personas analfabetas es muy superior al real.

OTRA MEDIDA PARA DETERMINAR EL PUNTO CENTRAL DE LA DISTRIBUCIÓN : LA MEDIANA

En el apartado anterior hemos podido comprobar que en las distribuciones campaniformes la media se sitúa en el centro de la distribución. Hemos podido ver igualmente que, cuando la distribución no es campaniforme, la media aparece desplazada con respecto al centro de la distribución.

Cuando la media no es una medida representativa de los valores de la variable se puede utilizar la mediana, que es otra medida distinta para determinar el punto central de la distribución. Para conocer el valor de la mediana se ordenan los datos de la variable, de forma creciente o decreciente; a continuación se busca el valor central que divide la distribución en dos mitades; a dicho valor se le denomina mediana. Cuando el número de valores de la variable es par, para obtener la mediana se debe calcular el promedio de los dos valores centrales de la distribución.



En la tabla siguiente podemos ver los datos correspondientes a las cantidades de cartón y papel que se recogieron en las distintas comunidades autónomas españolas en 2004 mediante procesos de recogida selectiva de basuras. Los datos reflejan, para cada una de las comunidades autónomas, la cantidad media, medida en kg., de papel y cartón producida por persona y año. Para obtener la mediana de estos datos, puesto que el número de comunidades autónomas es par, calculamos el promedio de los dos valores centrales de la distribución.

Comunidades autónomas	Papel y cartón (Kg.)
Extremadura	8,3
Ceuta y Melilla	9,3
Castilla y León	10
Andalucía	11
Castilla la Mancha	12,6
Valencia	13
Murcia	13,6
Galicia	14
Canarias	14,5
Cantabria	14,8
Aragón	15,8
Madrid	20,3
Cataluña	21,2
La Rioja	21,2
Asturias	22,8
Islas Baleares	25,7
Navarra	26
Euskadi	38,9

$$Me = \frac{14{,}5 + 14{,}8}{2} = 14{,}65$$

Fuente: Centro Regional de Estadística de Murcia. Anuario estadístico de la ciudad de Murcia. 2007.
http://www.carm.es/econet/anuario/a2007/anuario.html

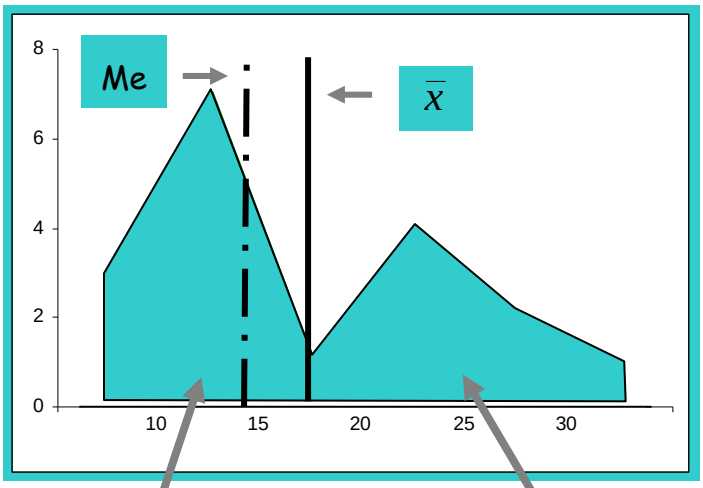
Además de la mediana, calcularemos ahora también la media, ya que el objetivo es mostrar la diferencia que existe entre los dos valores:

$$\frac{\sum_{i=1}^{n} x_i - \overline{x}}{n} = \frac{313}{18} = 17{,}39$$

Veamos ahora la diferencia entre las dos medidas:

	Media	Mediana
	17,39	14,65
Diferencia	2,74	

Comparados los resultados de las dos medidas –media y mediana- podemos concluir que la diferencia entre los valores de ambas es considerable: 2,7 Kg. de papel y cartón por habitante al año supone un aumento del 18,7% en relación a la mediana. Si miramos la gráfica del polígono de frecuencias en seguida comprenderemos por qué los valores de media y mediana son tan diferentes:



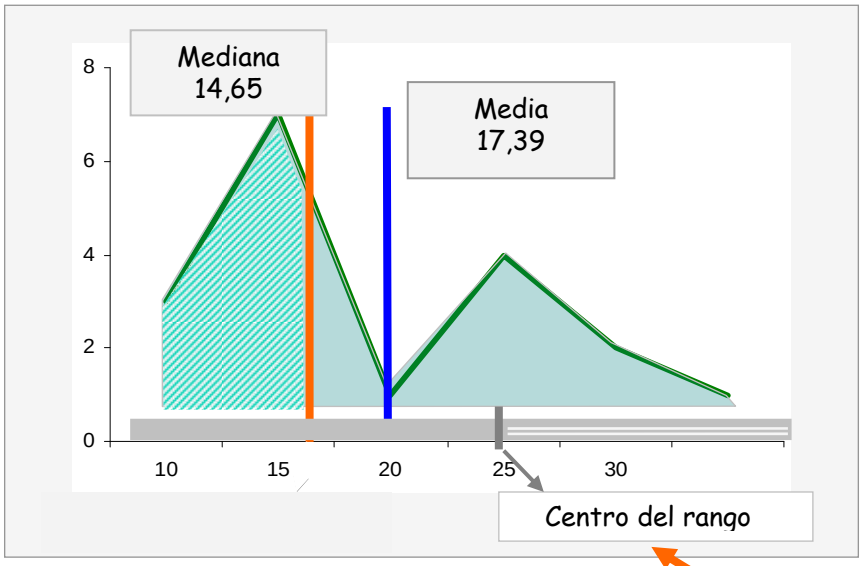
Grupo A: recogieron una cantidad inferior a 16 Kg.

Grupo B: recogieron una cantidad superior a 20 Kg.

Distribución bimodal

El polígono de frecuencias nos muestra una imagen reveladora del diferente comportamiento de las 18 comunidades autónomas en relación a la cantidad de papel y cartón que produjeron por persona y año en 2004. La gráfica nos permite observar también que existen dos grupos de comunidades. El primero, y más numeroso, está formado por 11 comunidades, en las cuales la cantidad de papel y cartón recogidos por persona no llegó a los 16 kg. El segundo de los grupos está formado por las 7 comunidades restantes; en ellas las cantidades de papel y cartón recogidas fueron muy superiores, alcanzándose en un caso los 38,9 Kg..

Resulta fácil comprender ahora que son los valores del segundo grupo de comunidades autónomas los responsables de que el resultado de la media sea sensiblemente superior al de la mediana. De hecho, el valor de la media, que se sitúa en medio de los dos subgrupos, no es un valor representativo de ninguno de ellos. 17,39 Kg. por persona y año es bastante más de lo que se recogió en ninguna de las primeras once comunidades y mucho menos también de lo que se recogió en las siete comunidades restantes.

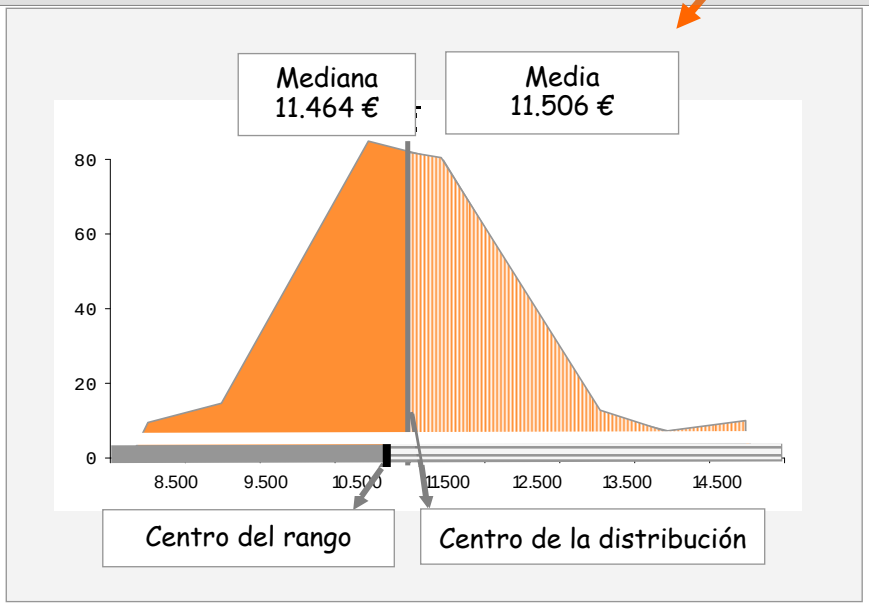


Cantidad de papel y cartón recogido en 2004 en las CCAA españolas. Distribución no campaniforme

Mediana, media y centro del rango: **valores muy diferentes**

Mediana, media y centro del rango: valores **muy similares**

Renta personal disponible. Municipios de Euskadi. 2001. Distribución campaniforme



En Estadística, la media aritmética y la mediana, junto con otras medidas, se clasifican en el grupo denominado medidas de tendencia central. Para ahora, ya es evidente la razón que justifica esta denominación, que no es otra que su *tendencia* a situarse en el centro de la distribución. Precisamente cuando esto no ocurre, es decir, cuando la media se aleja sensiblemente del sector central, deja de resultar un valor idóneo para representar a los datos.

EL PROMEDIO EN DISTRIBUCIONES CON VALORES EXTREMOS

El promedio es una medida muy sensible a la presencia de valores extremos, que pueden llegar a influir notablemente en su resultado. Para comprender esta idea hay que tener en cuenta que para calcular el promedio utilizamos todos los valores de la variable que estamos analizando y por ello, todos ellos tienen influencia en el resultado final. El problema surge cuando alguno de estos valores ejerce una influencia desmesurada, elevando o descendiendo notablemente el resultado del promedio.

La presencia de valores extremos es una situación habitual. En ocasiones, dentro de una serie de datos cuya distribución es campaniforme, existe algún dato especialmente alto o bajo con respecto a los demás. Esto es, por ejemplo, lo que ocurre con los datos del precio medio de la vivienda nueva en las principales ciudades aragonesas, en 2007. En la tabla adjunta podemos ver la gran diferencia que existe en el precio medio de la vivienda entre la ciudad más cara -Zaragoza- y la segunda de la lista -Jaca-. El precio de la vivienda en Zaragoza, más de 100.000 euros mayor que el de Jaca, constituye un valor extremo que eleva notablemente el valor del promedio.

	Ciudades	Precio	Incremento
1	Caspe	103.900	
2	Fraga	127.000	23.100
3	Barbastro	128.500	1.500
4	Alcañiz	129.700	1.200
5	Ejea de Los Caballeros	155.400	25.700
6	Calatayud	158.600	3.200
7	Teruel	170.600	12.000
8	Huesca	194.900	24.300
9	Jaca	197.600	2.700
10	Zaragoza	301.500	103.900

[1]La columna *incremento* nos permite ver la diferencia en el precio medio de la vivienda entre una ciudad y la siguiente más cara de la lista

[1] Precio medio, nominal, en euros, de la vivienda nueva en junio de 2007. Sociedad de Tasación. SA. http://web.st-tasacion.es/html/index.php. Última consulta 10-09-2008. *Valor nominal: En el valor se incorpora el IPC (unidad monetaria habitual a nivel de consumo). Valor Real: Valor en el supuesto de que el incremento del IPC hubiese sido 0 desde diciembre de 1985.*

Si calculamos el promedio con el dato de Zaragoza y sin él, comprobamos hasta qué punto influye este valor en el resultado: la diferencia llega prácticamente hasta los 15.000 euros.

	Sin Zaragoza	Con Zaragoza
Promedio	166.770	151.800

Diferencia: 14.970 €

A la vista de los resultados, es preciso concluir que cuando tenemos una serie de datos en la que existe algún valor especialmente alto o bajo tenemos que prestar especial atención a la influencia que estos valores pueden tener en el promedio.

Veremos ahora nuevos datos sobre el precio medio de la vivienda en 2007, pertenecientes también a la Comunidad autónoma de Aragón, pero esta vez sólo de la provincia de Zaragoza.[2] Se trata de los precios medios de 2007, a nivel comarcal. A diferencia de los datos anteriores, estos incluyen la vivienda nueva y la de segunda mano. El objetivo no es hacer ninguna comparación con los datos anteriores ni obtener conclusiones con respecto al precio de la vivienda. Se trata exclusivamente de mostrar que los valores extremos no siempre afectan de forma significativa el valor del promedio.

[2] Precio de la vivienda por metro cuadrado en las comarcas de Zaragoza. 2007. Fuente: Mercado Inmobiliario de Aragón 2007. Caja de Ahorros de la Inmaculada de Aragón.

	Comarca	Precio	Incremento
1	Bajo Martín	52.647	
2	Bajo Aragón-Caspe	78.960	26.313
3	Campo De Daroca	90.090	11.130
4	Sierra de Albarracín	92.584	2.494
5	Campo de Belchite	95.184	2.600
6	Campo de Borja	97.468	2.284
7	La Ribagorza	100.152	2.684
8	Matarraña	100.386	234
9	Sobrarbe	104.958	4.572
10	Andorra-Sierra de Arcos	105.840	882
11	Jiloca	107.334	1.494
12	Ribera Alta Del Ebro	108.864	1.530
13	Bajo Aragón	109.896	1.032
14	Valdejalón	112.800	2.904
15	Cinca Medio	121.869	9.069
16	Cinco Villas	122.400	531
17	Comunidad de Calatayud	123.172	772
18	Bajo Cinca	127.846	4.674
19	Gúdar-Javalambre	142.389	14.543
20	Somontano de Barbastro	145.376	2.987
21	Tarazona Y El Moncayo	153.510	8.134
22	La Jacetania	167.014	13.504
23	Ribera Baja Del Ebro	176.157	9.143
24	Los Monegros	178.830	2.673
25	Hoya de Huesca	181.764	2.934
26	Comunidad de Teruel	185.472	3.708
27	La Litera	186.400	928
28	Campo de Cariñena	186.944	544
29	Alto Gállego	192.975	6.031
30	D.C. Zaragoza	236.672	43.697

En este caso, también, es el dato de Zaragoza el que constituye un valor muy superior al del resto de la población. En la columna de la derecha de la tabla, en la que hemos calculado los incrementos de precio que se producen de cada comarca a la siguiente, podemos ver que el precio de Zaragoza supera en 43.697 euros al de la comarca anterior. Expresado en porcentajes, se puede decir que el precio medio en la comarca Zaragoza es un 22,6% superior al de la comarca del Alto Gállego.

Veamos ahora cuál es la influencia del valor de Zaragoza en el promedio:

	Con el dato de Zaragoza	Sin el dato de Zaragoza
Promedio	132.865,1	129.285,55
Diferencia	3.579,55	

En este caso también comprobamos que el valor de Zaragoza eleva el resultado del promedio en 3.579,55 euros. La cantidad no es despreciable, pero sí es muy inferior a los 15.000 euros que hemos obtenido anteriormente. Es evidente que, en este último caso, el valor de Zaragoza no es tan extremo como en el ejemplo anterior. Pero hay otro factor que explica la menor influencia en el promedio del elevado valor de Zaragoza: el tamaño de la población. Mientras la población de ciudades está integrada por 10 elementos, la población de comarcas cuenta con 30 elementos: el elevado valor de Zaragoza se reparte entre ellos de modo que se difumina su influencia en el promedio.

Imaginemos ahora que el precio medio de la vivienda en la comarca de Zaragoza es todavía mayor, por equipararlo con el dato extremo del precio medio de la vivienda en las ciudades. Imaginemos que en lugar de 236.672 es de 290.000 y veamos cómo repercute este incremento –ficticio- en el resultado del promedio:

	Con el dato de Zaragoza		Sin el dato de Zaragoza
Promedio	Ficticio	Real	
	134.643	132.865,1	129.285,5
Diferencia	1.779,9		3.579,6

Los resultados muestran que el dato ficticio que hemos creado incrementa la diferencia entre las dos comarcas más caras casi hasta 100.000 €. Pese a este notable aumento, el promedio se eleva sólo en 1.779 €

A la vista de los resultados, podemos concluir que, incluso elevando de forma notable el valor correspondiente a la comarca de Zaragoza, su influencia en el promedio sigue siendo discreta.

La influencia variable de los valores extremos en función del tamaño de la población, es decir, del número de valores que analizamos, se aprecia muy claramente en el caso de la renta personal disponible de los municipios de Euskadi en 2001. Uno de los 251 municipios tuvo un valor de renta muy superior al del resto. El valor de renta del municipio de Laukiz, con 20.627 €, constituye un valor extremo. La diferencia entre el valor de Laukiz y el valor del siguiente municipio en la lista, Lanestosa, es de 5.281 €.

Renta personal disponible de los municipios de Euskadi en 2001 (€)		
Valor mínimo	Lanestosa	7.192
Valor máximo	Laukiz	20.627
Segundo valor más alto	Leintz Gatzaga	15.346
	Incluido el valor extremo	Excluido el valor extremo
Promedio	10.514,47	10.474
Diferencia de 40,45 euros		

En la tabla superior podemos observar que la influencia del valor extremo de Laukiz en el promedio es mínima: la diferencia entre el promedio calculado con todos los valores y el promedio calculado excluyendo el dato de Laukiz se reduce a 40,45 €. Una vez más, podemos comprobar que en el caso de poblaciones con un gran número de elementos, la influencia de los valores extremos en el promedio es muy leve.

¿Qué podemos concluir, finalmente, sobre el modo de analizar una serie de datos que contiene valores extremos?

- En general, cuando el número de elementos de la población no es pequeño, la existencia de un elemento extremo no altera sustancialmente el valor del promedio.

- No hay ninguna norma o convención que permita determinar en qué momento el tamaño de una población deja de ser pequeño. Por esta razón la influencia de los valores extremos en el cálculo del promedio debe ser valorada en cada caso, es decir, siempre que detectemos su presencia.

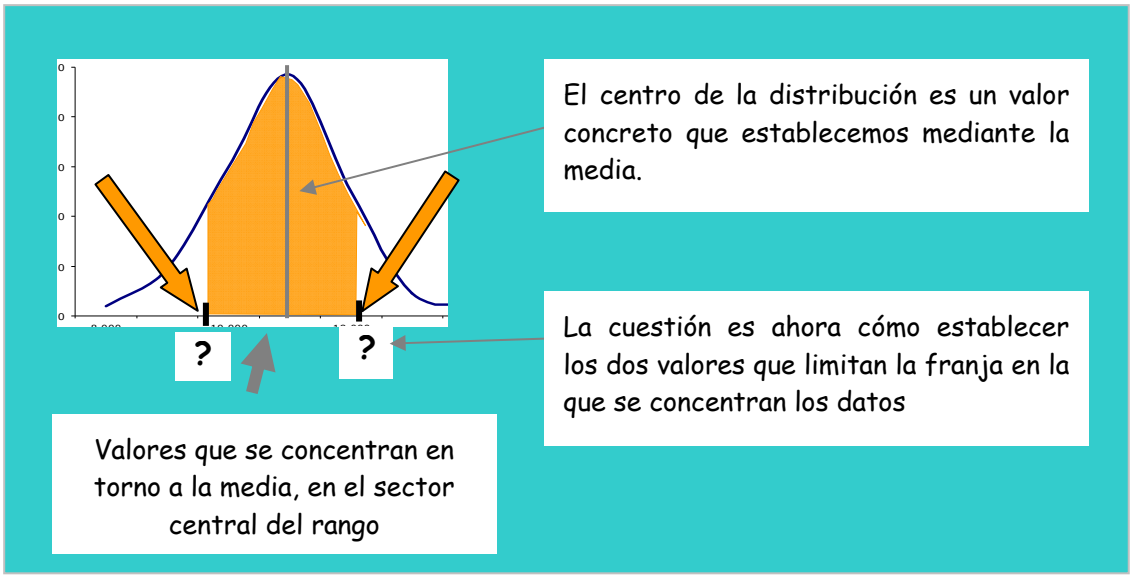
- Cuando consideremos que la influencia de algún valor extremo en el promedio es significativa, podemos tomar la opción de realizar el análisis excluyendo el o los valores extremos. Es una práctica habitual.

- En cualquier caso, los valores extremos, si no son errores o anomalías, en tanto que en realidad existen, no se pueden olvidar o simplemente apartar. Cuando excluimos los valores extremos del cálculo del promedio, no los estamos excluyendo del análisis estadístico. Lo que hacemos es calcular el punto central de la distribución sin elementos distorsionantes, pero en nuestro análisis y conclusiones les damos la importancia que tienen.

EL INTERVALO DE CONCENTRACIÓN DE LOS VALORES. LA DESVIACIÓN ESTANDAR

Introducción

Al hablar de las distribuciones campaniformes decíamos que la media es un valor que se sitúa en el centro del rango, y en el centro también del grupo mayoritario de valores.



De ese grupo de valores en torno a la media nos ocuparemos a continuación, es decir, del intervalo dentro del rango en el que se concentra una mayoría de los valores de la variable que analizamos. Para conocer dicho intervalo lo que hacemos es calcular los valores o puntos del rango que lo limitan.

Es importante comprender bien la lógica de los procedimientos de cálculo que explicaremos a continuación. Pero antes de eso, lo fundamental es tener muy claro por qué y con qué fin se utilizan estos procedimientos:

- El promedio y el intervalo que agrupa una mayoría de los datos de una variable se calculan cuando queremos conocer hasta qué punto, o en qué medida, son variables (diferentes entre sí) los valores de una variable.

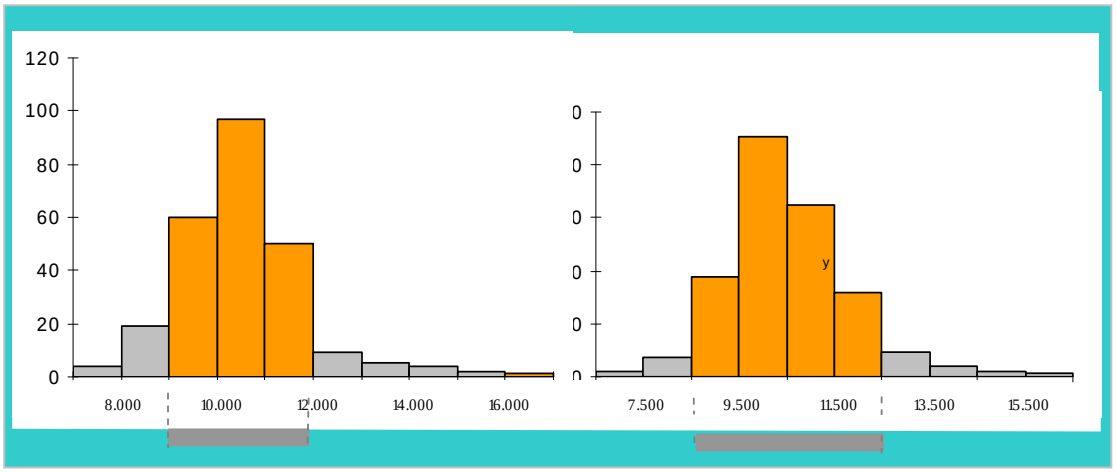
- Las variables son características o propiedades de los elementos de una

población. De una población compuesta por todas las empresas registradas, por ejemplo, en Guipúzcoa, una característica o variable podría ser el número de empleados que tiene cada una.

- Dado que los valores de las variables suelen ser distintos para cada uno de los elementos de la población, una forma de conocer, y de describir, su variabilidad es calcular en qué rango de valores se encuentra la mayor parte de los miembros de la población. Aplicado al ejemplo anterior, intentaríamos conocer cuáles son el número máximo y mínimo de trabajadores entre los que se encuentra la mayoría de las empresas.

PERO, es preciso recordar también que sólo podemos realizar este tipo de análisis si la distribución es campaniforme.

Al explicar los histogramas y polígonos de frecuencia hemos podido ver que este tipo de gráficos permite generar la imagen de cualquier distribución de valores y destacar en ella los intervalos que acumulan la mayor parte de los valores de una variable. En base a esta idea podríamos concluir que bastaría con utilizar el histograma de una distribución para calcular el intervalo que concentra una mayoría de los valores. La conclusión sería errónea puesto que la imagen que proporciona un histograma puede ser muy variable en función del número y de la amplitud de los intervalos que se elijan. Lo que necesitamos es un método que nos permita calcular de forma inequívoca el intervalo que buscamos.



Los dos histogramas que vemos en la imagen están elaborados en base a los mismos datos. Los dos contienen un número idéntico de intervalos, de igual amplitud. Lo que cambia entre ambos es el punto de inicio del histograma. El resultado es que la imagen que proporciona cada uno sobre el intervalo que acumula una mayoría de los

valores es distinta. En el gráfico de la izquierda la mayor parte de los valores de la variable se acumula entre 9.000 y 12.000; en el de la derecha la percepción de agrupamiento es menor y el intervalo se desarrolla entre 8.500 y 12.500. (Municipios de Euskadi. Renta personal disponible en 2001)

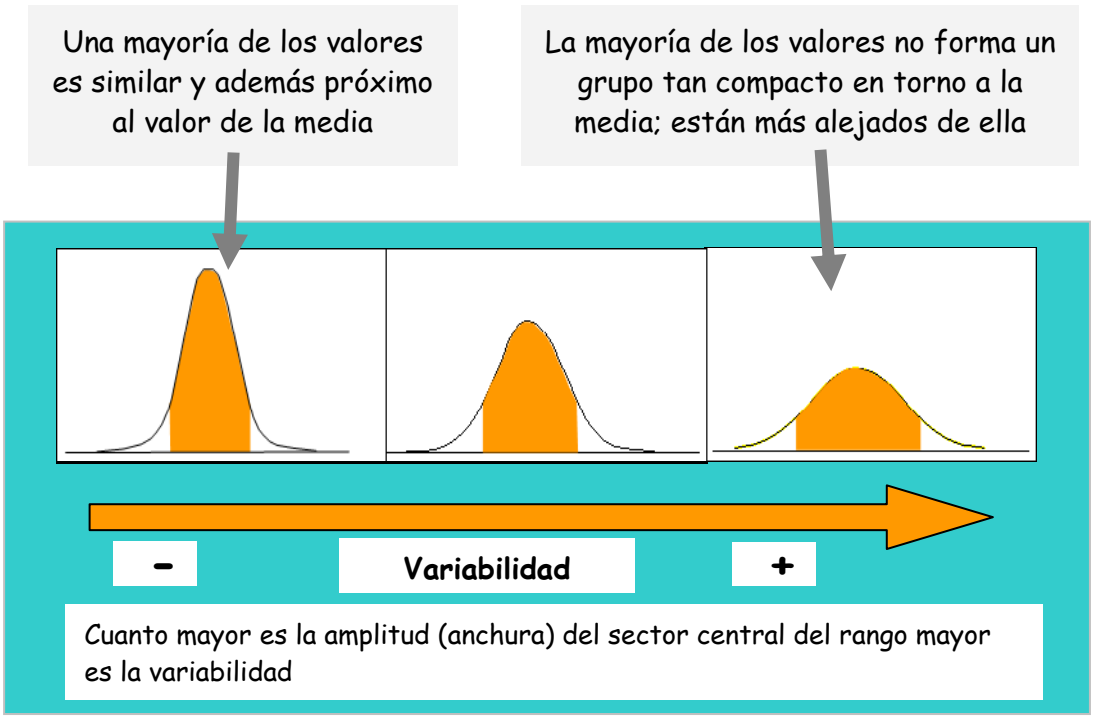
Para calcular el intervalo que agrupa una mayoría de los valores necesitamos un método que cumpla una serie de condiciones:

- Que su procedimiento de cálculo sea independiente de las decisiones que pueda tomar el usuario.
- Que, utilizados los mismos datos, proporcione siempre idénticos resultados.
- Que sea conocido, aceptado y utilizado por la comunidad científica.

La utilización de métodos estándar para el análisis estadístico permite no sólo que los resultados de nuestros estudios puedan ser comprendidos por todos aquellos que conocen los métodos sino que dichos resultados puedan ser comparados con los de otros estudios.

Antes de abordar la explicación sobre el procedimiento para calcular los límites del intervalo que agrupa la mayoría de los datos, es preciso introducir otra idea.

Aunque la concentración de valores en el sector central del rango es una característica de las distribuciones campaniformes, el grado de concentración puede ser muy variable. En algunas distribuciones una mayoría de los valores son muy semejantes y se encuentran muy próximos al valor de la media; en otras distribuciones el grupo mayoritario de valores no se sitúa tan cercano al valor central y conforma un intervalo más amplio dentro del rango.



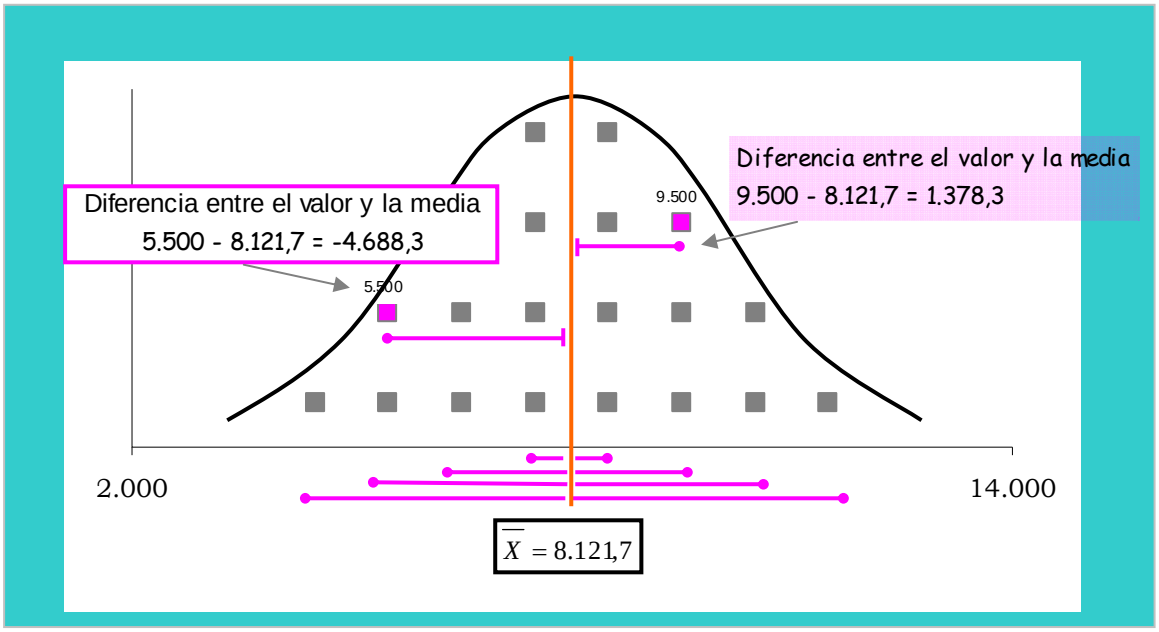
Las diferencias en el grado de concentración de los datos son en realidad diferencias en la variabilidad. De hecho, cuando intentamos determinar el sector del rango en el que se concentran buena parte de los valores de la variable, lo que hacemos en realidad es una valoración de la variabilidad que presenta esa mayoría de valores que se agrupa, en mayor o menor medida en torno a la media.

El concepto de desviación estándar y su procedimiento de cálculo

La desviación estándar es un procedimiento estadístico que se utiliza para determinar en qué medida se aleja la mayoría de los datos de la media. En el lenguaje estadístico, para referirnos a este alejamiento o distancia que existe entre los valores y la media, se utiliza el concepto de dispersión. Cuando los valores están alejados de la media, repartidos por todo el rango, se dice que están muy dispersos o que la dispersión es muy elevada. Se entiende así que la desviación estándar forme parte de un conjunto de herramientas que se conocen con el nombre genérico de medidas de dispersión.

Lo esencial, una vez más, es entender la idea o el principio que subyace al procedimiento de cálculo de la desviación estándar: la medida de dispersión de

los valores de la variable en torno a la media se basa en el cálculo de la distancia (diferencia) que existe entre los distintos valores y el valor de la media. A mayor lejanía, mayor dispersión y viceversa.



Los cuadrados grises representan los datos

Las barras representan la distancia o diferencia entre los valores individuales y el valor de la media

Siguiendo el principio mencionado, mediante la desviación estándar lo que se calcula es la diferencia entre cada uno de los valores de la variable y el valor de la media.

Este procedimiento nos proporciona un conjunto de valores y lo que a nosotros nos interesa es disponer de un valor global de la diferencia o distancia que hay entre los valores de la variable y la media. Para lograrlo lo que se hace es calcular un valor medio de las distancias de todos los valores a la media.

Para ganar claridad en la explicación expondremos el procedimiento de cálculo mediante un ejemplo. Utilizaremos para ello los datos del precio de la vivienda

en distintas ciudades de Euskadi en 2007[3]. En la tabla que mostramos a continuación podemos ver el listado de las 13 ciudades escogidas con sus correspondientes precios. Puesto que el objetivo es calcular la diferencia entre los valores individuales y el valor de la media, hemos dividido las ciudades en dos grupos, separando así las ciudades en las que el precio de la vivienda es inferior al de la media y las ciudades en las que es superior a ésta.

Valores de la variable: precio del m^2 en viviendas nuevas, €/m^2				
Valores inferiores a la media: diferencias negativas			Valores superiores a la media: diferencias positivas	
Basauri	- 2.120	Promedio 2.881,31	+ 2.916	Barakaldo
Hernani	- 2.255		+ 2.988	Gasteiz
Erandio	- 2.571		+ 3.097	Irun
Arrasate	- 2.577		+ 3.268	Bilbo
Portugalete	- 2.753		+ 3.272	Getxo
Santurtzi	- 2.781		+ 4.061	Donostia
Leioa	- 2.798			

A continuación calcularemos las diferencias individuales entre los valores y la media. Los valores que están a la izquierda del promedio tendrán diferencias negativas con respecto a éste; los que están a la derecha, tendrán diferencias positivas.

[3] Precio medio nominal, en euros por metro cuadrado, de la vivienda en junio de 2007. Sociedad de Tasación, SA. http://web.st-tasacion.es/html/index.php

Diferencias negativas		Diferencias positivas	
Ciudades	$x_i - \bar{X}$	$x_i - \bar{X}$	Ciudades
Basauri	-761,31	34,69	Barakaldo
Hernani	-626,31	106,69	Gasteiz
Erandio	-310,31	215,69	Irun
Arrasate	-304,31	386,69	Bilbo
Portugalete	-128,31	390,69	Getxo
Santurtzi	-100,31	1.179,69	Donostia
Leioa	-83,31		
$\sum x_i - \bar{X}$	-2.314,1	2.314,1	

Como decíamos anteriormente, el cálculo de las diferencias entre los valores de la variable y la media nos proporciona un nuevo listado de datos. Sin embargo, repetimos que lo que nos interesa a nosotros es disponer de un valor global, de una única medida de dispersión. Lo ideal sería poder calcular el valor medio de todas las diferencias individuales. Sería lo ideal pero no es posible porque si sumamos todas las diferencias el resultado es 0.

Diferencias negativas	+	Diferencias positivas	=	0
-2.314,1	+	2.314,1	=	0

La suma de las diferencias entre los valores de la variable y el promedio, en este caso, y en todos, es siempre cero, ya que se trata de una propiedad de la media aritmética:

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

Una vez que hemos comprendido la imposibilidad de calcular el promedio de las diferencias entre los valores de la variable y el promedio, nos resultará más sencillo entender el procedimiento estadísticos existentes para obtener un valor medio de las diferencias.

La estadística nos ofrece al menos dos alternativas para obtener un valor medio de la distancia de los valores de la variable a la media. Una de las alternativas es utilizar lo que se denomina *Desviación media.* Su cálculo consiste en eliminar el signo de las diferencias entre los valores y el promedio. Al eliminar el signo se suman las cantidades absolutas[4] y se calcula el promedio:

$$D\overline{x} = \frac{\sum_{i=1}^{n} x_i - \overline{x}}{n}$$

Aunque el cálculo de la desviación media es muy sencillo, en la práctica se usa muy poco como medida de desviación precisamente porque exige trabajar con valores absolutos.

Cálculo de la *Desviación media* aplicada a los datos sobre el precio medio nominal de la vivienda en 13 ciudades de Euskadi:

[4] El valor absoluto de un número real es su valor numérico, prescindiendo del signo.

$$\frac{\sum_{i=1}^{n} x_i - \bar{x}}{n} = \frac{4.628{,}31}{13} = 356{,}02$$

La segunda alternativa, y en la práctica la más utilizada, es la ***desviación estándar***. En este caso también el procedimiento se inicia mediante el cálculo de las diferencias entre los valores de la variable y el promedio. A partir de aquí, la solución para evitar que se compensen las diferencias positivas y las negativas consiste en elevar las diferencias al cuadrado. Desaparecen de este modo los valores negativos y es posible entonces calcular el promedio de las diferencias.

Para ilustrar el procedimiento continuaremos analizando los datos sobre el precio de la vivienda:

Ciudades de Euskadi	Valores de la variable	Diferencias entre los valores y el promedio	Cuadrado de las diferencias
	x_i	$(x - \overline{x})$	$(x - \overline{x})^2$
Basauri	2.120	-761,31	579.592,92
Hernani	2.255	-626,31	392.264,22
Erandio	2.571	-310,31	96.292,30
Arrasate	2.577	-304,31	92.604,58
Portugalete	2.753	-128,31	16.463,46
Santurtzi	2.781	-100,31	10.062,10
Leioa	2.798	-83,31	6.940,56
Barakaldo	2.916	34,69	1.203,40
Gasteiz	2.988	106,69	11.382,76
Irun	3.097	215,69	46.522,18
Bilbo	3.268	386,69	149.529,16
Getxo	3.272	390,69	152.638,68
Donostia	4.061	1.179,69	1.391.668,50
		$\sum_{i=1}^{n}(xi - \overline{x})^2$	2.947.164,77

Después de calcular las diferencias, elevarlas al cuadrado y realizar la suma, calculamos el promedio:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{2.974.164,77}{13} = 226.704,98\ €$$

Al hacer el promedio del cuadrado de las desviaciones para el ejemplo, nos damos cuenta de que obtenemos un resultado cuyo valor no guarda relación alguna con los valores de la variable. Mientras el precio por metro cuadrado de las viviendas estudiadas oscila entre 2.120 y 4.061 €, el valor de la *desviación estándar* se eleva hasta 226.704, 98 €. La razón es evidente: se trata de una cifra tan alta porque hemos elevado al cuadrado las diferencias entre los valores individuales y el valor de la media. Para compensar la elevación obtenemos la raíz cuadrada del promedio:

$$S = \sqrt{\frac{2.947.164,77}{13}} = 476,14$$

La última operación que hemos realizado, que nos proporciona un resultado cuya cifra es acorde al precio de la vivienda en las ciudades estudiadas, constituye el último paso en el cálculo de la desviación estándar, cuya fórmula queda como sigue:

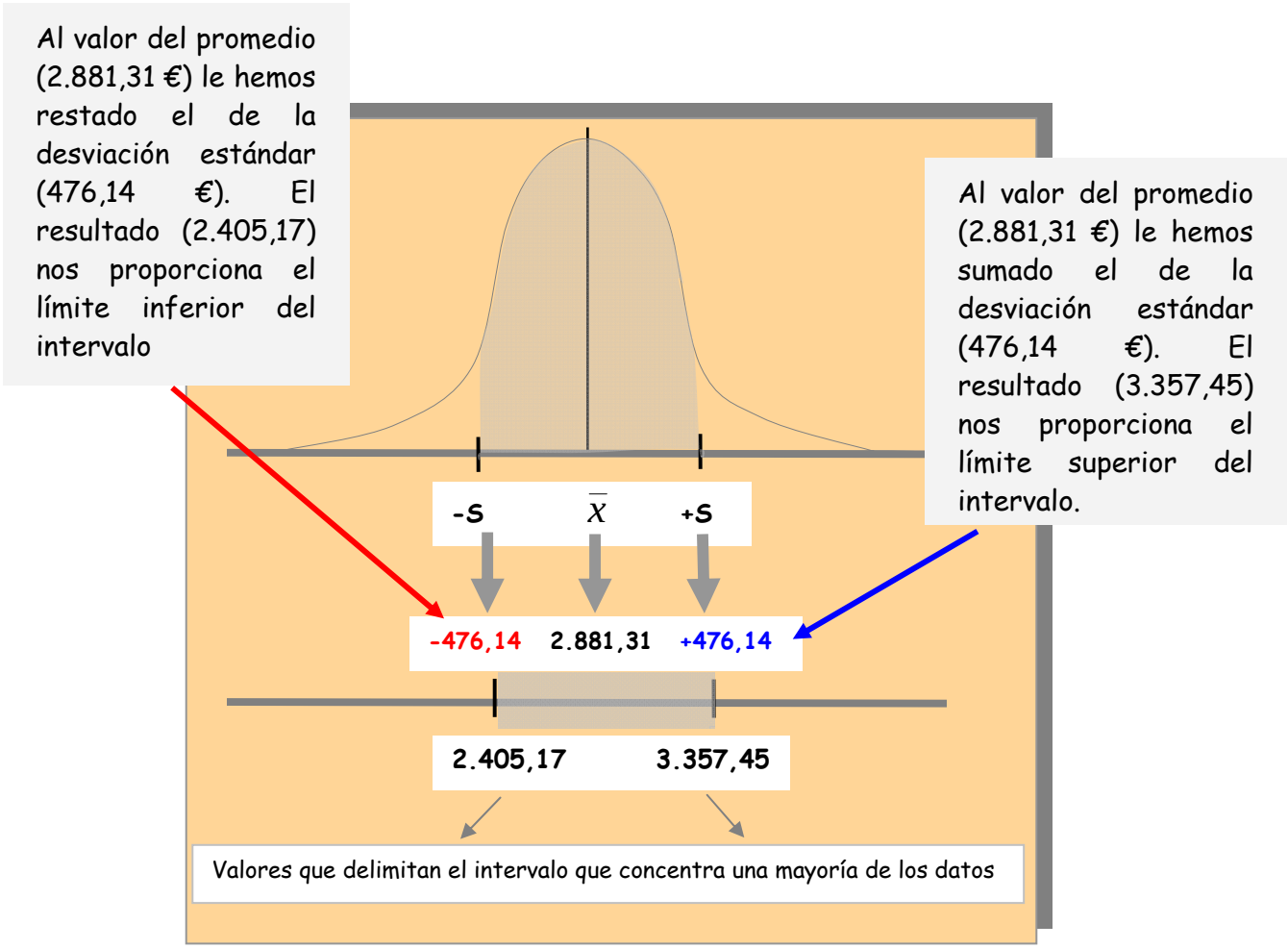
$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

La utilización de la desviación estándar

Al principio del apartado decíamos que nuestro objetivo era buscar el modo de calcular los dos puntos del rango entre los que se concentra una gran parte de los valores de la variable. Hemos dicho también que para calcular esos dos puntos se utiliza la ***Desviación estándar*** y hemos mostrado cómo se calcula. Ahora sólo nos falta explicar cómo se utiliza el valor de la *desviación estándar* para calcular el intervalo que concentra una mayoría de los valores de la variable.

Si aceptamos que la ***desviación estándar*** es una medida del grado en que los valores de la variable se alejan del promedio, nos resultará sencillo entender que una forma de calcular el intervalo que concentra una mayoría de los valores

consiste en sumar y restar al promedio el valor de la *desviación estándar*. Los dos valores resultantes de estas operaciones son los límites del intervalo que buscamos. Para ilustrar el procedimiento seguiremos utilizando los datos sobre el precio medio de la vivienda en trece ciudades de Euskadi:



En la ilustración aparecen reflejadas las operaciones que hemos realizado :

Una vez calculados los límites del intervalo estamos en situación de afirmar que en una mayoría de las ciudades analizadas el precio de la vivienda por metro cuadrado oscila entre los 2.405,17 y los 3.357,45 euros.

Todavía existe otra forma más de utilizar el valor de la desviación estándar para describir la variabilidad de los valores de una variable. En lugar de calcular los límites del intervalo que agrupa una mayoría de los valores, podemos mencionar directamente el resultado de la desviación estándar como medida

del alejamiento o dispersión de los valores con respecto a la media. Aplicado al caso de los precios de la vivienda, podríamos afirmar lo siguiente: la variabilidad con respecto a la media del precio medio de la vivienda en trece ciudades de Euskadi es de 476,14 €. Esto significa que los valores de la variable se dispersan una media de 476,14€ con respecto a la media aritmética.

A estas alturas resulta evidente que, cuanto mayor es el resultado de la desviación estándar, mayor es el grado de variabilidad o de dispersión de los valores con respecto al promedio. Ahora bien, no olvidemos que, para calificar la dispersión de alta o de baja, no debemos fijarnos en el valor en bruto de la desviación sino compararlo con el valor del promedio. Una dispersión, supongamos, de 500 € referida a los precios de ropa de abrigo cuyo promedio es de 800 €, es, sin lugar a duda, un valor muy elevado. El mismo valor de dispersión referido al precio de vehículos cuyo promedio es de 24.000 €, es, por el contrario, muy bajo.

De todos modos sólo a veces es posible calificar de alto o bajo un valor de desviación estándar. En la mayoría de situaciones no disponemos de referentes que nos permitan añadir un calificativo de grado al valor de la desviación. Fijémonos en los resultados obtenidos para los datos del precio medio de la vivienda en las ciudades de Euskadi. ¿Cómo podríamos calificar el resultado de la desviación (476,14€) en relación al valor del promedio (2.881,31 €)?

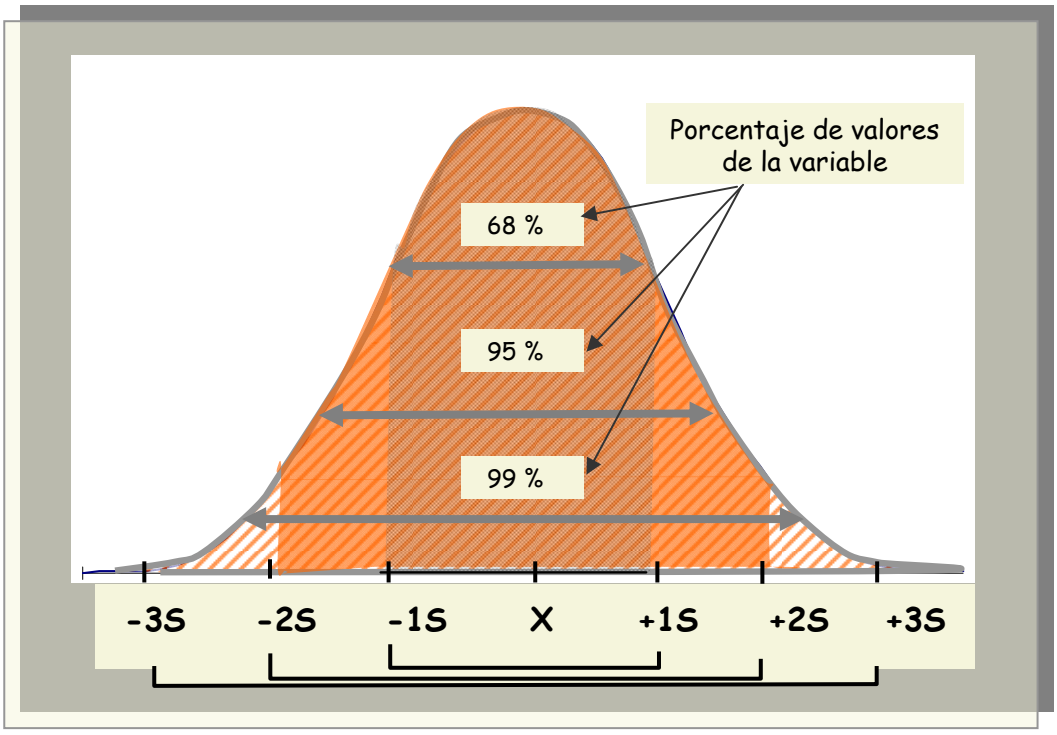
Las profesoras Neer y Lopetegui proporcionan comentarios interesantes en relación a la valoración de la desviación estándar:

> Solemos preguntarnos cuándo una desviación es grande y cuándo pequeña. Por un lado la interpretación de grande o pequeña depende de la variable que estemos analizando, por ejemplo una desviación de 5 será grande si estamos hablando de la dispersión de la edad en que los niños aprenden a leer y será pequeña si nos estamos refiriendo a la dispersión del salario de los habitantes de La Plata. Por otro lado grande o pequeño tienen un significado relativo, es más grande o más pequeño que el encontrado para algún otro grupo o algún otro test. (Neer, Lopetegui, 2003)

Tal como hemos explicado, a la hora de elaborar una descripción de la variabilidad de los datos podemos utilizar el valor de la desviación estándar para calcular el intervalo de agrupamiento de valores o como un valor medio de variabilidad o dispersión. Podemos también utilizar ambos modos en distintos momentos de la explicación. En cualquier caso, sea cual fuere la forma

de expresión que adoptemos, es importante que incluyamos en nuestro trabajo una pequeña tabla con los resultados de todas las medidas y técnicas que hemos utilizado y, dentro de ella junto al promedio, el resultado de la desviación estándar.

Cuando en los resultados de un estudio proporcionamos el valor de la desviación estandar o los límites del intervalo $\overline{X} \pm S$, estamos afirmando, aunque no sea de forma explícita, que una gran mayoría de los valores de la variable estudiada se encuentra dentro de los límites de dicho intervalo. El posible lector de nuestro trabajo sabe que esto debe ser así. La razón es que en todas las distribuciones campaniformes se cumple la condición de que esa mayoría de valores de la que hablamos se sitúa en el intervalo de la media +/- la desviación. De hecho, las distribuciones campaniformes cumplen las siguientes condiciones:



- El 68% de los valores de la variable se encuentra dentro del intervalo formado por la media +/- la desviación.

- El 95% de los valores de la variable se encuentra dentro del intervalo formado por la media +/- dos veces el valor de la desviación.

- El 99% de los valores de la variable se encuentra dentro del intervalo formado por la media +/- tres veces el valor de la desviación.

Otros usos de la desviación estándar

El valor de la desviación estándar se utiliza frecuentemente también como medida para identificar valores de la variable especialmente alejados del promedio, es decir, notablemente más altos o más bajos que aquel. El objetivo suele ser destacar la existencia de algunos elementos de la población cuyo comportamiento respecto a la variable estudiada se pueda considerar especial. En este caso no estaríamos hablando de los valores extremos (outliers) a los que nos referíamos en el tema tres.

El procedimiento al que nos referimos consiste en sumar y restar al promedio el valor de la ***desviación estándar*** multiplicado por dos o por tres. Mediante la suma y la resta obtenemos los límites de un intervalo; estos límites se utilizan para separar los valores que se consideran especialmente altos o bajos. Los límites del intervalo marcan los puntos del rango a partir de los cuales se sitúan los valores destacables.

Una vez explicado el procedimiento es preciso aclarar la forma correcta de aludir al mismo. Cuando se utiliza la doble desviación estándar hablamos del intervalo que surge del promedio más/menos ***dos unidades de desviación estándar*** y lo expresamos como sigue: $\boldsymbol{\bar{X}} \pm 2\boldsymbol{S}$. Siguiendo esta lógica, cuando utilizamos la triple desviación hablamos del intervalo que surge del promedio más/menos ***tres unidades de desviación estándar*** y lo expresamos así: $\boldsymbol{\bar{X}} \pm 3\boldsymbol{S}$

Calcularemos ahora los intervalos que surgen del promedio más/menos ***dos y tres unidades de desviación estándar*** para los datos referentes al precio medio de la vivienda en las trece ciudades de Euskadi:

Promedio	2.281,31
Desviación estándar	476,14

	Promedio	Unidades de desviación estándar								Resultado
		1		2		3				
$\overline{X} \pm 2S$	2.881,31	- 476,14	-	476,14	=					1.929,04
		+ 476,14	+	476,14	=					3.833,58
$\overline{X} \pm 3S$	2.881,31	- 476,14	-	476,14	-	476,14	=			1.452,90
		+ 476,14	+	476,14	+	476,14	=			4.309,71

Límites del intervalo del promedio ± 2 uds. de Ds

1.929,04 — 3.833,58

Límites del intervalo del promedio ± 3 uds. de Ds

1.452,9 — 4.309,71

- En ninguna de las ciudades analizadas el precio medio de la vivienda supere los límites del intervalo $\overline{X} \pm 3S$.

- En una de las ciudades analizadas, En Donosti, el precio medio de la vivienda supera el limite superior del intervalo $\overline{X} \pm 2S$.

Desde el principio hemos visto que el precio medio de la vivienda en Donosti, 4.061€/m², es muy superior al del resto de las ciudades analizadas. Utilizar el

intervalo $\overline{X} \pm 2S$ puede ser una forma idónea de describir hasta qué punto es elevado dicho precio.

Son muchos los estudios en los que se quiere particularizar la situación de los valores situados en los extremos del rango y que utilizan para ello los límites del intervalo $\overline{X} \pm 2S$ o $\overline{X} \pm 3S$: los elementos de la población con valores por encima o por debajo tendrán una consideración especial.

En un informe de septiembre de 2008, el Servicio Meteorológico Nacional mexicano, para destacar el elevado número de tormentas intensas ocurridas de enero a agosto de 2008, se expresaba en los siguientes términos:

> Durante el mes de agosto de 2008 se registraron a nivel nacional un total de **126 tormentas intensas** superando el promedio de 85.75 tormentas en el mes. El record de agosto de 2008, representa un valor superior al promedio más una desviación estándar. (CONAGUA, 2008)

En este caso, los autores del estudio han utilizado el intervalo del promedio más/menos una unidad de desviación estándar. Aunque es más frecuente que se utilicen los intervalos de dos o tres unidades de desviación, cuando el valor de la desviación es elevado, el intervalo del promedio más/menos una unidad de desviación también es muy amplio y se usa entonces como límite para detectar valores especialmente elevados.

En el Atlas de la industria en la Comunidad de Madrid, podemos ver un ejemplo concreto en el que el límite definido por el promedio más dos unidades de desviación estándar se ha aplicado para detectar los municipios especializados en actividades industriales concretas:

> CÁLCULO DE LOS VALORES DE ESPECIALIZACIÓN
>
> **El nivel de especialización en las diferentes actividades de la industria manufacturera se ha calculado mediante el índice de Nelson. El índice se basa en las propiedades de la desviación típica como medida de dispersión de los valores de una distribución, para discriminar aquellos que sobrepasan anormalmente determinados umbrales. Utilizando los sectores urbanos como unidad de referencia espacial, se obtiene en primer lugar el porcentaje de cada una de las once actividades industriales. A continuación se calculan para cada una la media aritmética y desviación típica. Los valores límite para considerar una zona como especializada se definen como el valor medio más una, dos, tres o más desviaciones típicas. En nuestro caso hemos elegido como umbral**

> mínimo el valor medio más dos desviaciones típicas. Puede ocurrir que algunas zonas no sobrepasen los valores límite establecidos, y se consideren que no están especializados en ninguna de las actividades. También pueden darse zonas especializadas en dos o más actividades. En este caso se ha asignado la actividad de mayor grado de especialización.
>
> Debe tenerse en cuenta que esta forma de cálculo de especialización funcional es relativa, en el sentido de que no da una medida de la importancia o peso total de una actividad, sino la importancia en relación al resto de las zonas consideradas. (Comunidad de Madrid, 2007)

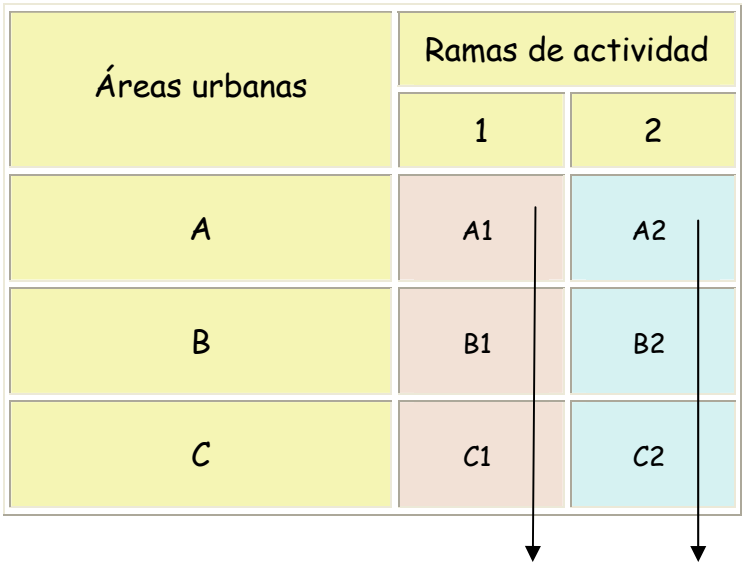
Tal como explica el texto, el punto de partida ha sido el cálculo, para cada área urbana de la Comunidad de Madrid de la importancia, medida en el porcentaje que representa cada una de las once ramas industriales con respecto al total de la industria en dicha área. Este cálculo produce una estructura de datos similar a la que mostramos a continuación:

Áreas urbanas	Ramas de actividad industrial										
	1	2	3	4	5	6	7	8	9	10	11
A	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
B	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11
C	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11

En la primera línea tendríamos los porcentajes de cada una de las actividades industriales en el área urbana A. La suma de los porcentajes en las 11 ramas de actividad (de A1 a A11), constituye el 100%.

Si lo miráramos por columnas, en la número 1 estarían los porcentajes de la rama industrial 1 de todas las áreas urbanas de la Comunidad de Madrid. La suma de los porcentajes de la actividad industrial nº 1 en todas las áreas urbanas (A1, B1, C1, etc.) constituye el 100%.

Partiendo de los porcentajes obtenidos, los autores del atlas calcularon la media y la desviación estándar de los datos por columnas, es decir, de los porcentajes por rama de actividad industrial.



Se calcula el promedio y la desviación estándar de los valores, en porcentajes, para cada una de las ramas de actividad industrial: A1, B1, C1, etc.

Una vez obtenidos estos datos pudieron calcular el valor del promedio más dos unidades de desviación estándar. El valor definido sirvió a modo de umbral a partir del cual se podía considerar que un área urbana estaba especializada en determinada rama de actividad industrial. Se pueden considerar especializadas en una rama concreta de actividad industrial aquellas áreas urbanas en las que el porcentaje de dicha rama de actividad supere el valor establecido por el promedio más dos unidades de desviación estándar.

La zona A se considera especializada en la rama de actividad 1 cuando su porcentaje de actividad industrial en dicha rama sea superior al promedio (de A1, B1, C1, etc.) más dos unidades de desviación estándar

En un estudio sobre la evolución de la precipitación anual en Andalucía, su autor utiliza las unidades de desviación estándar, sumadas o restadas al promedio, para generar una clasificación del grado de humedad de una serie de años en distintas regiones pluviométricas de Andalucía.

Como vemos en la tabla adjunta, para clasificar los años con una humedad superior a la media utiliza los límites que marcan el promedio más una, dos o tres unidades de desviación estándar.

- Los años cuya precipitación supera el límite del promedio más dos unidades de desviación estándar forman la clase de años extremadamente húmedos.
- El año se considera hiperhúmedo si la precipitación recogida supera el límite formado por el promedio más tres unidades de desviación estándar.
- En el extremo opuesto, para clasificar los años especialmente secos ha creado un único umbral, definido por el promedio menos una unidad de desviación estándar. (Castillo Requena, 2000)

AÑO MUY SECO	R año < $\overline{X} - S$
AÑO SECO	R año < $\overline{X} - 1/4\,S$ y > $\overline{X} - S$
AÑO NORMAL/SECO	R año < $\overline{X}$ y > $\overline{X} - S$
AÑO NORMAL/HÚMEDO	R año < $\overline{X}$ y > $\overline{X} + S$
AÑO HÚMEDO	R año < $\overline{X}$ y > $\overline{X} + S$
AÑO MUY HÚMEDO	R año < $\overline{X} + 1/4\,S$ y > $\overline{X} + S$
AÑO EXTREMADAMENTE HÚMEDO	R año< $\overline{X} + S$ y > $\overline{X} + 2S$
AÑO HIPERHÚMEDO	R año < $\overline{X} + 3S$

R: Media de la región pluviométrica

Para terminar con las explicaciones sobre la desviación estándar haremos un resumen de las características fundamentales de esta técnica:

- La desviación estándar (o desviación típica) es una herramienta para obtener una medida de la variabilidad de una serie de datos. Dentro de las herramientas estadísticas se incluye en la categoría de técnicas destinadas a la medición de la dispersión de los datos.
- La desviación estándar se puede utilizar como complemento al promedio, con el fin de proporcionar una medida de dispersión de los datos. Se puede utilizar también para calcular los límites del intervalo que concentra una mayoría de los valores. En cualquier caso, el valor de la desviación estándar es un complemento obligado del promedio.
- Como medida de dispersión el valor de la desviación estándar se interpreta en relación al valor del promedio.
- La desviación estándar se expresa en las mismas unidades de medida que la variable que analizamos. (Si la variable se refiere al precio en euros de la vivienda, la desviación estándar nos da un valor de la dispersión, también en euros)

El coeficiente de variación

Como hemos visto, la desviación estándar mide la desviación de los valores de la variable en las mismas unidades de la variable. En el ejemplo que veíamos sobre el precio del metro cuadrado de las viviendas en las distintas ciudades de Euskadi, hemos obtenido un precio medio de 2.281,31 euros por metro cuadrado. El resultado de la desviación estándar -476,14- también está expresado en euros por metro cuadrado. Esta característica de la desviación estándar resulta muy cómoda a la hora de interpretar la variabilidad de una distribución. Sin embargo no resulta útil cuando queremos comparar el grado de variabilidad de dos o más distribuciones. Veremos mediante un ejemplo cuál es el problema.

En la tabla siguiente podemos ver el precio por metro cuadrado de vivienda, correspondiente al año 2007, en los principales municipios de dos comunidades autónomas.

Ciudades de Euskadi	Precio de la vivienda €/m^2	Ciudades de Galicia	Precio de la vivienda €/m^2
Basauri	2.120	Monforte de Lemos	898
Hernani	2.255	O Barco de Valdeorras	1.080
Erandio	2.571	Ames	1.206
Arrasate	2.577	Ferrol	1.391
Portugalete	2.753	Pontevedra	1.477
Santurtzi	2.781	Viveiro	1.526
Leioa	2.798	Lugo	1.540
Barakaldo	2.916	Ourense	1.742
Gasteiz	2.988	Santiago de Compostela	1.786
Irun	3.097	Coruña	1.999
Bilbo	3.268		
Getxo	3.272		
Donostia	4.061		

Puesto que los datos de la tabla están ordenados es fácil darse cuenta de que la variabilidad entre los municipios en las dos comunidades es importante. En las dos comunidades autónomas el precio de la vivienda en las ciudades más caras es el doble, o casi el doble, del precio en las ciudades más baratas. Ahora bien: ¿en cuál de las dos comunidades autónomas es mayor la variabilidad?. Con el fin de intentar dar respuesta a la pregunta calcularemos los promedios y las desviaciones de los valores de la variable en cada comunidad.

	Euskadi	Galicia
Promedio	2.881,31	1.513
Desviación estándar	476,14	340,51

Las diferencias entre las dos comunidades son importantes, fundamentalmente en sus promedios. El País Vasco es la comunidad con el promedio más elevado

y también con la mayor desviación estándar. La cuestión es ahora decidir si con estos datos podemos saber en cuál de las dos comunidades es mayor la variación entre municipios.

A primera vista podríamos pensar que la mayor desviación estándar se corresponde con la mayor variabilidad y que es el País Vasco la comunidad con mayor variación interna de los precios; tal conclusión sería un error. Recordemos que la variabilidad medida mediante la desviación estándar se corresponde con la distancia o diferencia media que existe entre el promedio y los valores de la variable. Cuanto más elevados sean los valores de la variable más elevado será, en general, el valor del promedio y el de la desviación estándar.

Lo que ocurre en el ejemplo que hemos puesto es que los precios en el País Vasco son mucho más elevados que los de Galicia. Consecuentemente, los valores del promedio y de la desviación en el País Vasco tenderán a ser mayores que los de Galicia, incluso aunque la dispersión de los datos sea menor.

La conclusión parece clara: el valor de la desviación estándar no sirve para comparar el grado de variación interna de dos o más series de datos si los valores de dichas series no son de magnitud muy similar.

Existe, sin embargo, una herramienta que sí nos permite hacer comparaciones entre la variabilidad de diferentes conjuntos de datos, aunque los valores de unos sean mucho mayores o menores que los de los otros. La herramienta a la que nos referimos es el coeficiente de variación. Se trata de una herramienta de fácil manejo mediante la cual se calcula la magnitud de la desviación estándar en relación a la magnitud del promedio.

$$cv = \frac{S}{\overline{X}} \cdot 100$$

El coeficiente de variación se calcula mediante una simple división entre el promedio y la desviación estándar. El resultado se multiplica por 100 y se obtiene así el porcentaje que representa la desviación típica con respecto a la media

Podemos calcular ahora los coeficientes de variación para los valores del precio de la vivienda en las dos comunidades y comprobaremos que no es el País Vasco la comunidad en la que existe mayor variación de los precios entre municipios:

$$CV_{PaísVasco} = \frac{S}{\bar{x}} = \frac{476{,}14}{2.881{,}31} = 0{,}1652 \rightarrow 16{,}52\%$$

$$CV_{Galicia} = \frac{S}{\bar{x}} = \frac{340{,}51}{1.513} = 0{,}2251 \rightarrow 22{,}51\%$$

Los resultados del coeficiente de variación muestran que, para los datos del País Vasco, el valor de la desviación estándar supone un 16,52% del valor de la media. En el caso de Galicia, el valor de la desviación estándar es mayor y supone concretamente un 22,51% del valor de la media.

A la vista de los resultados del coeficiente de variación podemos afirmar que, de las dos comunidades autónomas, es la de Euskadi la que presenta una menor dispersión o variabilidad de los valores en torno a la media. Es cierto que, de media, los pisos son más baratos en Galicia pero también es cierto que el precio presenta mayores variaciones entre las ciudades gallegas que entre las vascas.

En el estudio que hemos mencionado anteriormente sobre la eevolución de la precipitación anual de Andalucía, su autor utiliza también el coeficiente de variación para destacar la elevada variabilidad interanual de las precipitaciones en Andalucía, en comparación con el resto de España.

	Precipitaciones anuales		
	Media	Desviación estándar	Coeficiente de variación
Cuenca del Guadalquivir	600,47	180,45	30,05
Cuenca Sur	537,82	165,72	30,81
Unidad Central	605,26	127,55	21,07
Fachada Mediterránea	527,69	116,60	22,10
España peninsular	667,09	119,25	17,88

Pese a tener un valor de desviación superior al de la Cuenca Sur, el coeficiente de variación es menor

La variabilidad de la precipitación en la España peninsular es notablemente inferior a la de las regiones andaluzas

Castillo Requena, J.M. (2000)