Tema 6:

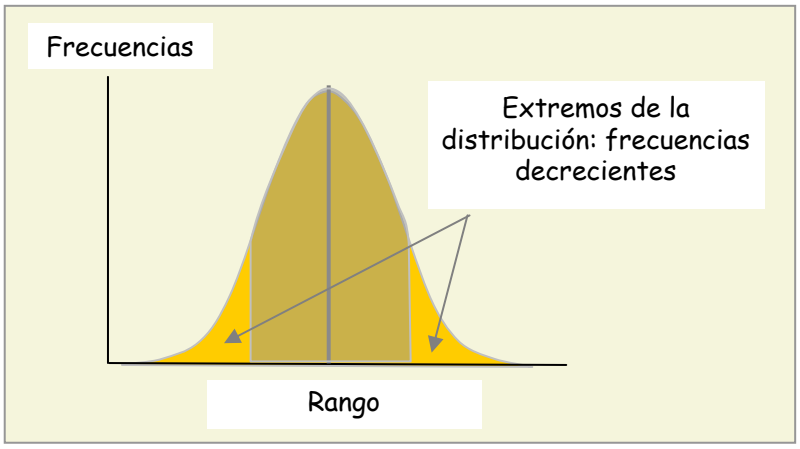
Análisis de los extremos de la distribución. La simetría

Tema 6:

Análisis de los extremos de la distribución. La simetría

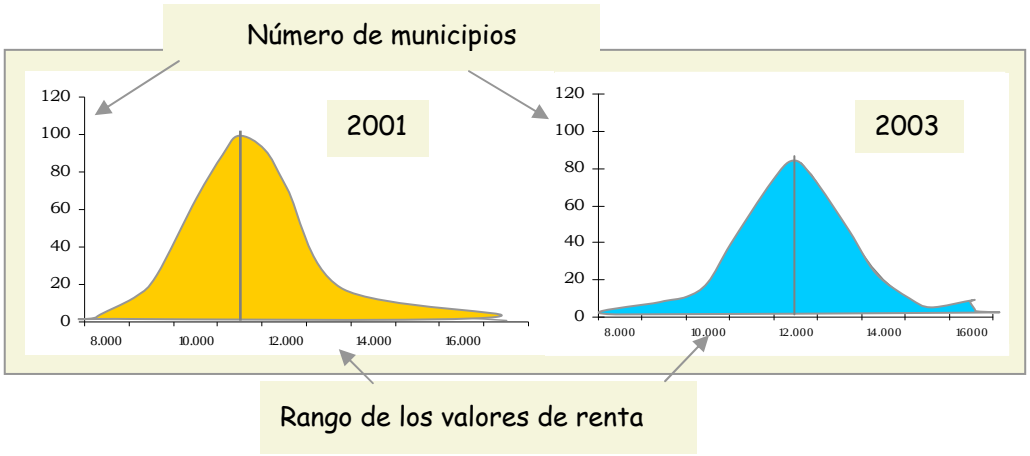
En el tema anterior hemos explicado el procedimiento para analizar el sector central de la distribución. Hemos visto que para realizar dicho análisis se utiliza la media aritmética junto con la desviación estándar. En este tema nos centraremos en el procedimiento para estudiar los lados de la distribución, para lo cual se utiliza el análisis de simetría.

Ya hemos visto, en repetidas ocasiones, que en las distribuciones campaniformes los extremos de la distribución representan al conjunto de los valores más altos y más bajos de la variable. Se trata de dos grupos porcentualmente minoritarios cuyas bajas frecuencias producen un descenso más o menos brusco de la curva de frecuencias:



Aunque esta es la forma prototípica de las distribuciones campaniformes -su representación ideal- muchas de las variables que se ajustan a este tipo de distribución presentan ciertas variantes o diferencias con respecto a ella.

EL CONCEPTO DE SIMETRÍA



Los polígonos de frecuencias de la ilustración se corresponden con la distribuciones de la variable[1] Renta personal disponible en los municipios de Euskadi. El polígono de frecuencias de la izquierda está elaborado con los datos de 2001 y el de la derecha con los datos de 2003.

Podemos observar que las dos distribuciones son de tipo campaniforme: ambas tienen un sector central que agrupa una mayoría de los datos; desde el centro las frecuencias van disminuyendo progresivamente a medida que desciende el número de municipios con rentas particularmente altas o bajas.

Las dos distribuciones son de tipo campaniforme pero presentan entre ambas diferencias de forma evidentes.

- En la gráfica de la izquierda la forma de la distribución a cada uno de los lados del eje central es diferente. El lado derecho es más ancho que el izquierdo y, sobre todo, en el tramo final de la curva se produce un alargamiento pronunciado de la misma que no está ni en su lado izquierdo ni en la distribución de los datos de 2003. Debido a la diferencia de forma entre los dos lados de la distribución, decimos que esta distribución es asimétrica.

[1] Ver el listado original de valores en el apartado de ANEXOS

- En la gráfica de la derecha los dos lados de la distribución son muy similares y, por ello, decimos que es simétrica.

Para entender el concepto de simetría acudiremos al diccionario:

> **Simetría: Arreglo equilibrado de partes de una figura en lados opuestos de un punto, línea, o plano. Los tipos más comunes incluyen la simetría con respecto a un punto, simetría con respecto a una línea y simetría rotacional.**[2]
>
> ***Geom.*** **Correspondencia exacta en la disposición regular de las partes o puntos de un cuerpo o figura con relación a un centro, un eje o un plano**[3]**.**

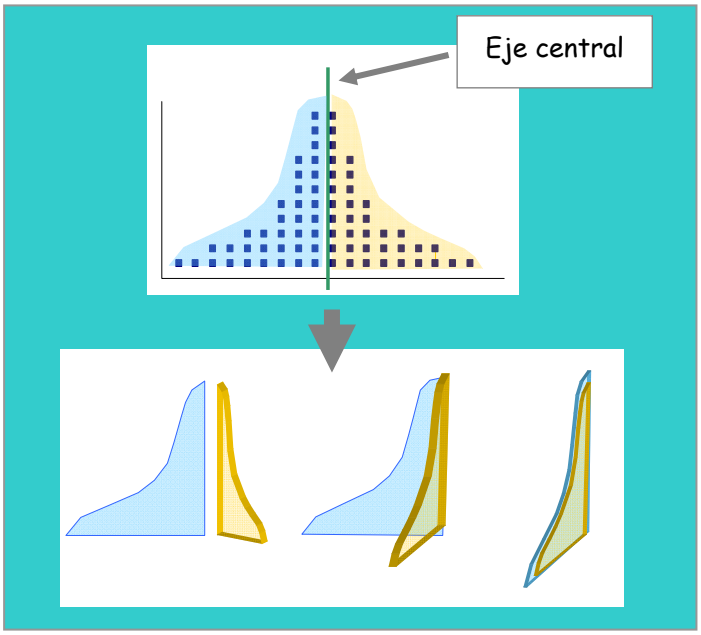
Vemos que se trata de un concepto perteneciente al ámbito de la geometría, que se utiliza en el análisis de datos y que resulta una valiosa herramienta para poner de relieve características significativas de las distribuciones.

ANÁLISIS GRÁFICO DE LA SIMETRÍA

Mediante la ilustración de la página anterior hemos visto a qué se refiere el concepto de ***simetría*** cuando hablamos de una distribución de valores de la variable. Podremos afirmar que una distribución es simétrica si al dividirla en dos a partir del eje de simetría, la forma de los dos lados es igual.

[2] http://www.mathematicsdictionary.com/spanish/vmd/full/s/symmetry.htm

[3] Diccionario de la Real Academia Española de la lengua. RAE

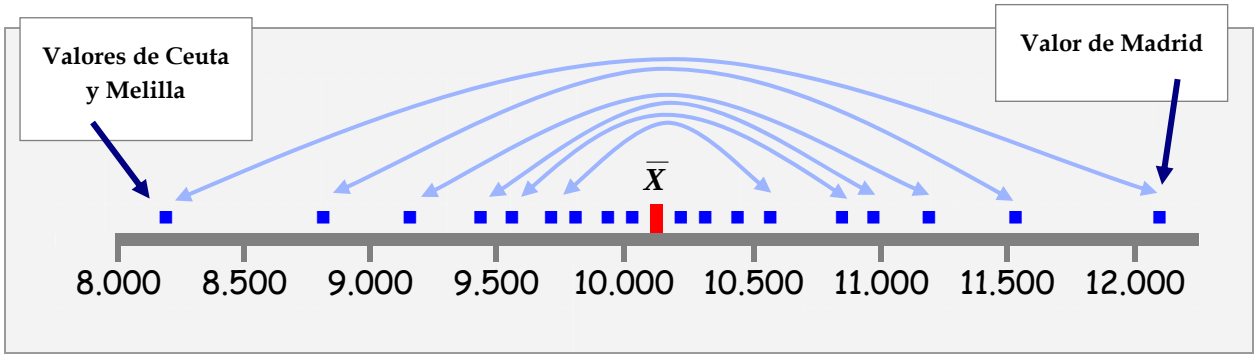


Cuando la distribución es simétrica, al doblarla por la mitad y juntar los dos lados de la distribución, ambos coinciden.

La similitud entre ambas partes nos permite identificar el carácter simétrico o asimétrico de una distribución. A este respecto, lo fundamental es tener claro lo que significa la simetría, o la asimetría, con respecto a un conjunto de datos y de qué modo podemos valorarla o cuantificarla.

El hecho de que los dos lados de una distribución sean iguales o similares significa que el número de valores a un lado y al otro del promedio es similar y se encuentran también a distancias similares de aquel. Esto significa a su vez que la dispersión de los datos a ambos lados del promedio es también similar.

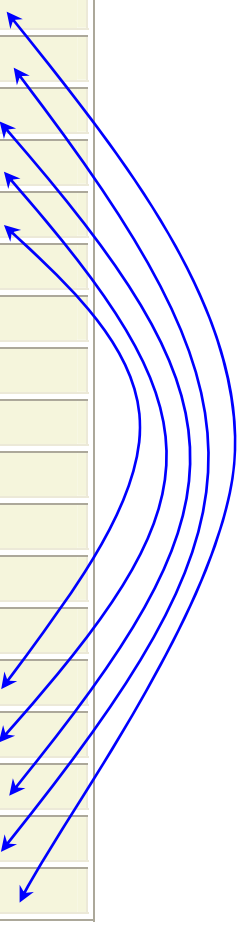
Para verlo más claro, representaremos la distribución de un conjunto hipotético de datos sobre una recta que representa el rango. Podemos suponer que los datos se corresponden con la cantidad media de dinero que gastaron un grupo de personas en cada una de las comunidades autónomas del estado español, durante 2006.



La marca roja en la recta representa el promedio, cuyo valor es de 10.150,98 euros. Las marcas azules representan el valor del gasto medio por persona en cada comunidad. Las líneas azules unen los valores equidistantes del promedio a uno y otro lado del mismo.

En la tabla siguiente podemos la serie original de datos (fictícia). En ella también hemos enlazado mediante líneas azules algunas de las parejas de variables cuya distancia al promedio es igual.

Comunidades autónomas	Gasto €	Diferencia con respecto a la media
Ceuta y Melilla	8.190,13	-1.960,85
Extremadura	8.565,99	-1.359,99
Castilla - La Mancha	8.945,31	-980,67
Canarias	9.257,72	-740,26
La Rioja	9.320,97	-563,22
Murcia	9.391,17	-438,06
Andalucía	9.554,63	-312,90
Cantabria	9.668,21	-187,74
Castilla y León	10.000,27	-62,58
Galicia	10.083,93	62,58
Asturias	10.662,51	187,74
Aragón	10.684,19	312,90
Comunidad Valenciana	10.720,13	438,06
Islas Balears	11.208,57	563,22
Navarra	11.542,70	740,26
País Vasco	11.855,74	980,67
Cataluña	11.994,94	1.359,99
Madrid	12.111,83	1.960,85



Si observamos la ilustración precedente y la tabla de valores de la variable podemos comprobar que los que gastaron más que la media y los que gastaron menos lo hicieron en la misma medida: vgr. el grupo de la Comunidad de Madrid gastó 600,86 euros más que la media y el grupo de las ciudades de

Ceuta y Melilla gastó 600,86 euros menos que la media. Si comparamos ahora Cataluña con Extremadura, veremos que los grupos gastaron 379,32 euros más que la media el primero y menos que la media el segundo.

Si hacemos la comprobación con todos los valores de la tabla obtendremos 9 parejas de comunidades. Los dos valores de cada pareja se sitúan a la misma distancia del promedio, uno por encima y el otro por debajo.

Parejas de comunidades equidistantes		Distancia al promedio
Ceuta y Melilla	Madrid	± 1.960,85
Extremadura	Cataluña	± 1.359,99
Castilla - La Mancha	País Vasco	± 980,67
Canarias	Navarra	± 740,26
La Rioja	Islas Baleares	± 563,22
Murcia	Comunidad Valenciana	± 438,06
Andalucía	Aragón	± 312,90
Cantabria	Asturias	± 187,74
Castilla y León	Galicia	± 62,58

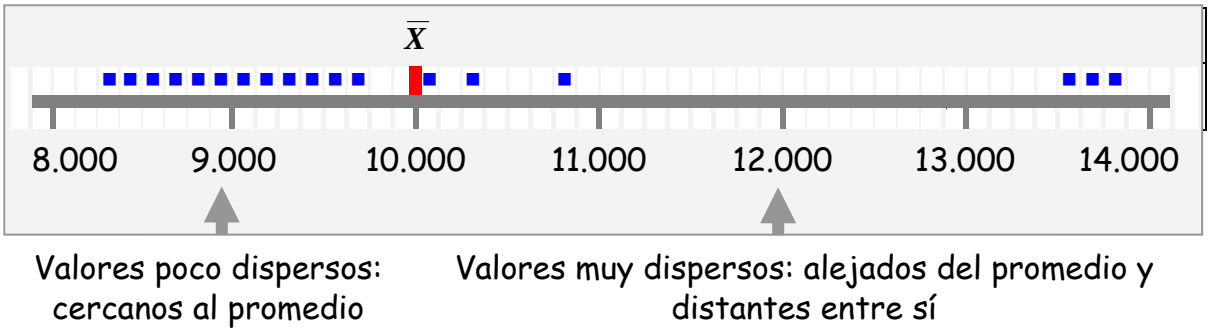
Finalmente, las conclusiones que se pueden obtener sobre la distribución de valores que hemos creado son las siguientes:

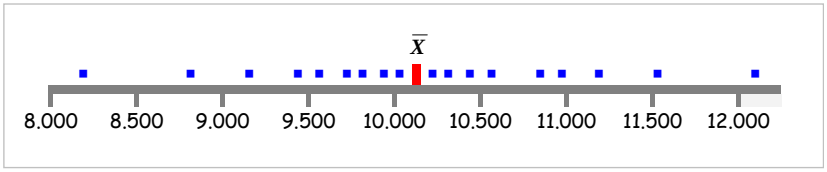
- Hay el mismo número de valores a la izquierda y a la derecha del promedio
- Los valores a ambos lados del promedio presentan la misma dispersión con respecto a éste.
- Se trata de una distribución totalmente simétrica.

Ahora bien, ¿qué significado tiene la simetría con respecto a la variable que estamos estudiando?

- Que no hay un grupo que destaque sobre el resto en su comportamiento con respecto al gasto. Quienes gastaron más y quienes gastaron menos estuvieron igual de cerca o de lejos del grupo mayoritario en torno al promedio.

Para comprender mejor el significado de la simetría el siguiente paso será plantearnos cómo sería una distribución asimétrica. Para ello utilizaremos también datos ficticios, relacionados con el mismo tema, pero modificando las cantidades del gasto asignadas a los distintos grupos y comunidades. (Promedio 10.051,61) (Ver tabla de datos en la última página del tema)





Véase la diferencia con la distribución del ejemplo anterior:

Un vistazo rápido basta para comprobar que la distribución en este caso es radicalmente diferente:

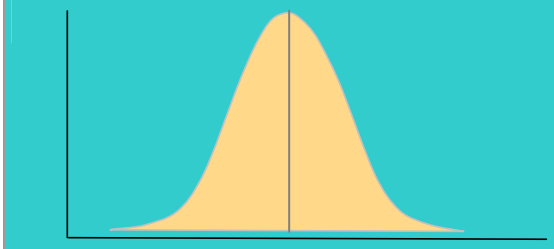
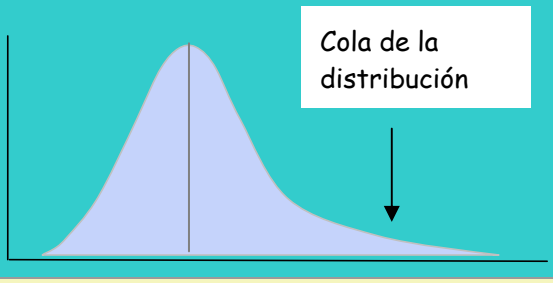
- 12 de los 18 valores se acumulan a la izquierda del promedio, muy cercanos a éste.
- A la derecha del promedio encontramos sólo un tercio de los valores. Tres de ellos se sitúan en el extremo derecho del rango, a mucha distancia del resto.
- El promedio, 10.051,61 euros, se sitúa muy desplazado hacia la izquierda dentro del rango.
- La distribución es claramente asimétrica: la dispersión a ambos lados del promedio es muy diferente. A su izquierda se concentra un grupo de valores, próximos a él y próximos también entre sí. A la derecha del

¿Qué significa en este caso la asimetría con respecto a la variable que estamos estudiando? ¿Cómo podemos interpretar el hecho de que la distribución sea asimétrica?:

Ya no podemos hablar, como en el ejemplo anterior, de un comportamiento muy homogéneo de los grupos con respecto al gasto. En aquel, veíamos a la mayoría de los grupos con gastos muy similares, próximos al promedio, y el resto de los grupos con gastos ligeramente inferiores o superiores.

En este caso, dentro del grupo de los 18 elementos que constituyen la población existe un subgrupo de 3 elementos que ha tenido un gasto notablemente superior al de los demás. No existe, sin embargo, ningún subgrupo que destaque por haber tenido un gasto especialmente inferior al del resto.

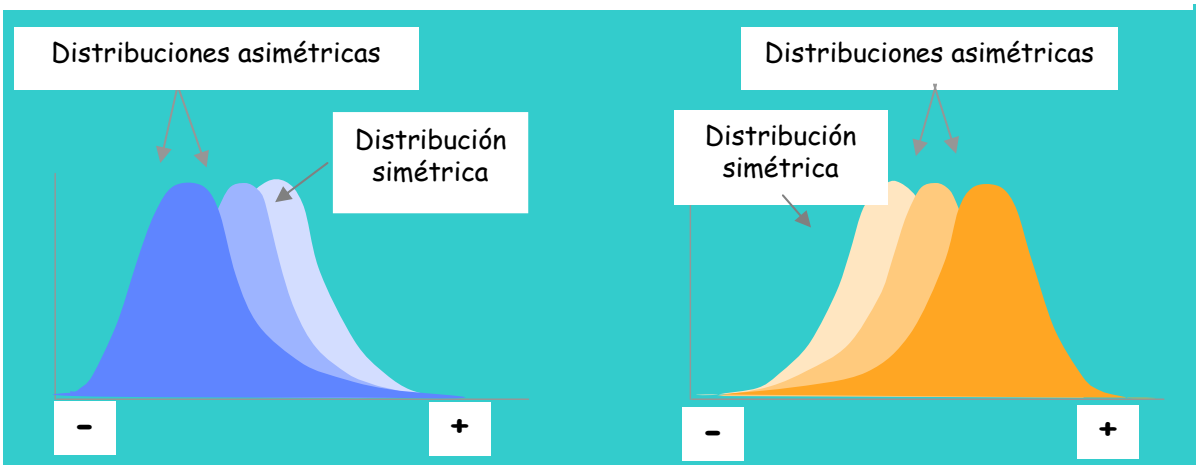
Si representáramos las distribuciones de frecuencias de los dos ejemplos que hemos utilizado, obtendríamos gráficas muy similares a las que veremos a continuación:

Gráfica de una distribución simétrica	Gráfica de una distribución asimétrica
	
Los dos lados de la distribución son iguales y, por tanto, simétricos	Debido a la presencia de valores mucho más elevados que la mayoría, el lado derecho de la gráfica se alarga, formando lo que se denomina cola de la distribución.

En el ejemplo de asimetría que hemos utilizado, la distribución se alarga hacia la derecha porque los elementos con fuerte dispersión se corresponden con valores elevados, esto es, con valores que se ubican en el extremo derecho del rango. Cuando esto ocurre, hablamos de que la asimetría positiva.

Lo mismo puede ocurrir con respecto al lado izquierdo de la distribución. En este caso los elementos con mayor dispersión que el resto se corresponderían con los valores más bajos, es decir, con aquellos que se ubican en el extremo izquierdo del rango de la variable. En este caso hablamos de asimetría negativa.

Representaciones de distribuciones asimétricas de distinto signo e intensidad



Asimetría positiva

El alargamiento se produce del lado de los valores que son mayores del promedio (+)

Asimetría negativa

El alargamiento se produce del lado de los valores que son menores del promedio (-)

El alargamiento de uno de los lados de la distribución puede ser grande o pequeño, dependiendo también del grado de dispersión de los valores de la variable.

Hemos visto que las distribuciones son asimétricas cuando la dispersión de los valores a uno y otro lado del promedio es distinta. Hemos visto igualmente que cuando hablamos de la asimetría de una distribución nos estamos refiriendo a una característica del conjunto de datos, a un aspecto significativo de la realidad que es preciso analizar y describir. Veremos ahora qué aporta el análisis de simetría a la descripción de una distribución y, por último, cómo calcular una medida de simetría de las distribuciones.

QUÉ APORTA EL ANÁLISIS DE SIMETRÍA

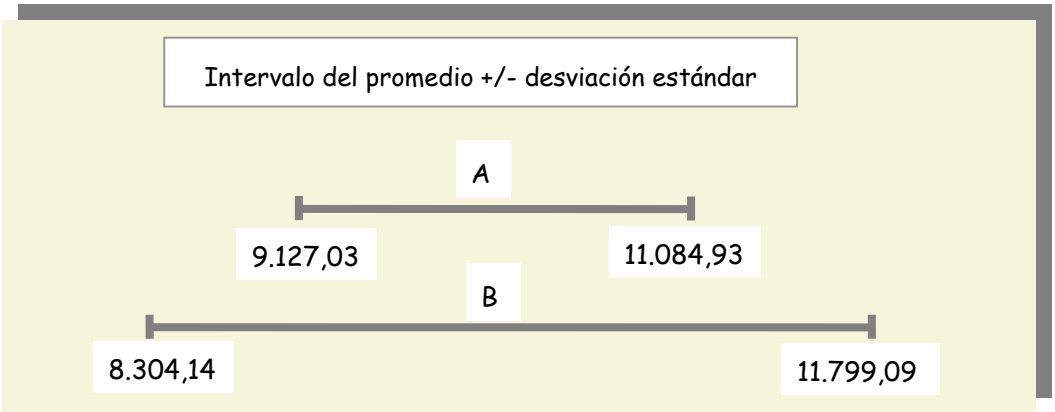
Con el fin de que se entienda mejor la necesidad de calcular el grado de simetría de las distribuciones y lo que esta medida nos proporciona veremos, mediante un ejemplo, las características de la distribución que se explican mediante el promedio y la desviación estándar. Esto nos servirá para evidenciar los rasgos de la distribución que dichas medidas no explican, rasgos que se pueden descubrir mediante el análisis de simetría. Utilizaremos los mismos datos que hemos usado anteriormente como ejemplos de distribuciones simétricas y asimétricas. Recordemos que se trata de datos ficticios, creados a partir de la

modificación de los datos pertenecientes a la encuesta de presupuestos familiares del INE en 2006. (Ver tabla de datos en la última página del tema)

Gasto total medio por persona. €. Comunidades autónomas.				
	Distribución simétrica **A**		Distribución asimétrica **B**	
Promedio	10.150,98		10.051,61	
Desviación estándar	933,95		1.747,47	
Promedio ± desviación est.	9.217,03	11.084,93	8.304,14	11.799,09
$\overline{X} \pm S$ amplitud del intervalo	1.868		3.495	
Valor mayor y menor	8.190,13	12.111,83	8.400,13	13.836,83
Rango	3.921,7		5.436,7	

Si atendemos al promedio de las dos distribuciones que figuran en la tabla, las dos pueden parecer similares; la diferencia entre los promedios de ambas apenas supera los 100 €. La conclusión será diferente si nos fijamos en los valores de la desviación estándar; la distribución B tiene un valor de desviación estándar que casi duplica al de la distribución A.

La diferencia en la variabilidad de las dos distribuciones es evidente. El intervalo formado por el promedio +/- la desviación estándar nos da una idea clara:



Las medidas obtenidas (promedio, desviación estándar, rango) han puesto en evidencia la mayor variabilidad de la distribución A con respecto a la B. Sin embargo, no nos dicen nada sobre las características de los datos que generan la mayor variabilidad de la distribución B. Nosotros, que hemos analizado y revisado los datos una y otra vez, sabemos que la mayor variabilidad de la distribución B se genera debido al gasto particularmente elevado de tres comunidades autónomas. El análisis y descripción de la variable que hemos estudiado resultaría ambiguo, y podría además generar interpretaciones equivocadas, si no incluimos en el mismo una medida que de cuenta de la asimetría de la distribución o, lo que es lo mismo, de la presencia de elementos de la población con valores de gasto muy por encima del promedio.

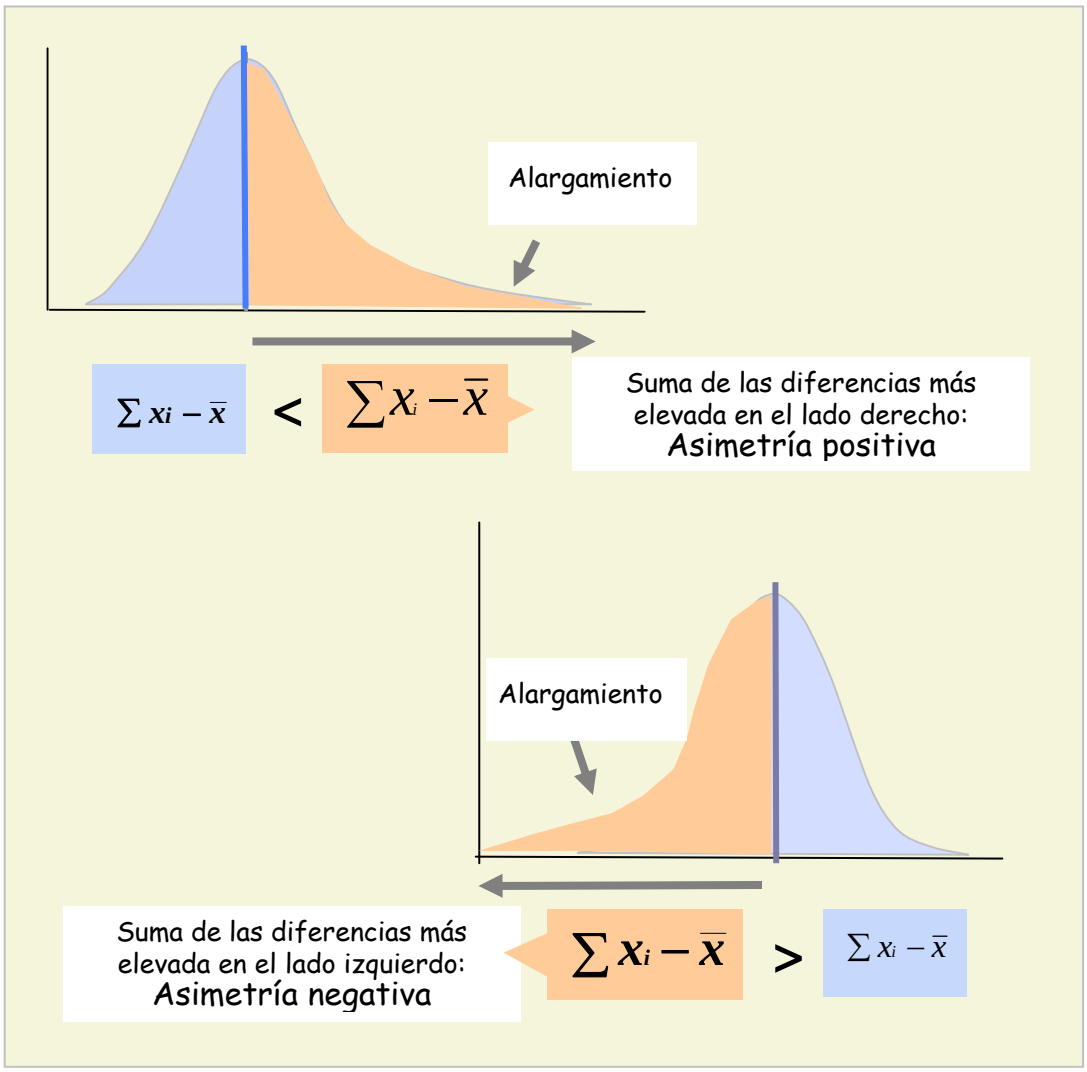
CUANTIFICACIÓN DE LA SIMETRÍA

El análisis gráfico de una distribución nos permite comprobar la existencia de asimetría en la distribución de los valores de la variable. Un vistazo al histograma o al polígono de frecuencias basta para descubrir el alargamiento de la gráfica a uno de los lados del rango. Sin embargo, a la hora de describir la asimetría de un conjunto de datos resulta mucho más cómodo disponer de una medida numérica.

La Estadística nos proporciona herramientas diferentes para obtener una medida de la simetría de las distribuciones. Una de ellas es el llamado coeficiente de asimetría de Fisher que, además de ser muy utilizada, resulta muy sencilla en su cálculo.

$$G_1 = \frac{\sum (x_i - \bar{x})^3}{N \cdot S^3}$$

El coeficiente de asimetría de Fisher se basa, una vez más, en el cálculo de las diferencias entre los valores de la variable y el promedio. En esta ocasión, las diferencias se elevan al cubo de modo que se mantiene el signo de las mismas; los valores por debajo del promedio proporcionan resultados negativos y los valores por encima del promedio dan resultados positivos. Consecuentemente, el resultado del coeficiente puede ser positivo o negativo:



- Cuando las diferencias de los valores a la izquierda de la media son superiores a las diferencias de los valores a su derecha, el resultado del coeficiente es negativo. Hablamos en ese caso de asimetría negativa que, gráficamente, se refleja en un alargamiento de la curva en el lado izquierdo de la distribución.

- Cuando las diferencias de los valores a la derecha de la media son superiores a las diferencias de los valores a su izquierda, el resultado del coeficiente es positivo. Hablamos en ese caso de asimetría positiva que, gráficamente, se refleja en un alargamiento de la curva en el lado derecho de la distribución

- Cuando las diferencias de los valores a ambos lados del promedio son iguales o muy similares, el resultado del coeficiente es próximo a cero. En este caso decimos que la distribución es simétrica.

Calcularemos ahora los coeficientes de asimetría de las dos distribuciones correspondientes al gasto medio total por persona en las comunidades autónomas del estado español. Reproduciremos nuevamente la tabla con los valores del promedio, desviación, etc.

Gasto total medio por persona. €. Comunidades autónomas.				
	Distribución simétrica A		Distribución asimétrica B	
Promedio	10.150,98		10.051,61	
Si miramos exclusivamente el promedio pensaremos que las dos distribuciones son muy similares				
Desviación estándar	933,95		1.747,47	
El valor de la desviación estándar nos pone en la pista de las diferencias de variabilidad entre las dos distribuciones				
Promedio ± desviación est.	9.217,03	11.084,93	8.304,14	11.799,09
$\overline{X} \pm S$ amplitud del intervalo	1.868		3.495	
Lógicamente, la amplitud del intervalo es muy superior en el caso de la distribución asimétrica				
Valor mayor y menor	8.190,13	12.111,83	8.400,13	13.836,83
Rango	3.921,7		5.436,7	
Coeficiente de asimetría	0		1,1	
	Simétrica		Asimétrica positiva	
Finalmente, es el coeficiente de asimetría el que nos aclara que es la presencia de algunos valores elevados en el extremo de la distribución la responsable de la diferencia de variabilidad entre las dos distribuciones. Esto significaría que, si exceptuamos las tres comunidades con gastos especialmente elevados, la variabilidad de los dos grupos de datos sería bastante similar.				

Gasto total medio por persona. €. Comunidades autónomas			
Diferencias negativas con respecto a la media $(x_i - \overline{x})^3$		Diferencias positivas con respecto a la media $(x_i - \overline{x})^3$	
Ceuta y Melilla	-5367188169	Cada uno de los resultados se ha obtenido restando al promedio el valor del gasto en la comunidad. El resultado se ha elevado al cubo.	
Extremadura	-4581326379		
Castilla - La Mancha	-3622817076		
Canarias	-2808293712		
Rioja (La)	-2126037535		
Murcia (Región de	-1564329795		
Andalucía	-1111451743		
Cantabria	-755684628		
Castilla y León	-485309700,9	Balears (Illes)	2694322,836
Galicia	-288608211,2	Navarra	261100906,6
Asturias (Principado de)	-153861409	País Vasco	40677415776
Aragón	-69350544,29	Cataluña	45275986328
Comunidad Valenciana	-1362094,039	Madrid (Comunidad de)	50074079137
-22.935.620.996,429		136.291.276.470,44	

Las diferencias positivas con respecto a la media son superiores a las diferencias negativas: la distribución es asimétrica positiva.

Gasto total medio por persona. €. Comunidades autónomas		
	Distribución simétrica A	Distribución asimétrica B
Ceuta y Melilla (Ciudades Aut.)	8.190,13	8.400,13
Extremadura	8.790,99	8.490,13
Castilla - La Mancha	9.170,31	8.615,13
Canarias	9.410,72	8.740,13
Rioja (La)	9.587,77	8.865,13
Murcia (Región de	9.712,93	8.990,13
Andalucía	9.838,09	9.115,13
Cantabria	9.963,25	9.240,13
Castilla y León	10.088,41	9.365,13
Galicia	10.213,57	9.490,13
Asturias (Principado de)	10.338,73	9.615,13
Aragón	10.463,89	9.740,13
Comunidad Valenciana	10.589,05	10.040,13
Balears (Illes)	10.714,21	10.290,13
Navarra (Comunidad Foral de)	10.891,24	10.790,13
País Vasco	11.131,65	**13.590,13**
Cataluña	11.510,97	**13.715,13**
Madrid (Comunidad de)	12.111,83	**13.836,83**

Valores responsables del incremento de la asimetría de B con respecto a A

La diferencia fundamental entre las dos tablas de datos reside en los tres últimos valores que, en el caso de la distribución asimétrica, significan que el gasto en tres de las comunidades autónomas ha sido sensiblemente superior al del resto de las comunidades. Este comportamiento particularizado de las tres comunidades con mayor gasto sólo se puede identificar utilizando el coeficiente de asimetría.