CAPITULO 2.- ANALISIS DE UNA VARIABLE .

2.1 Introducción.

En este capítulo se darán un conjunto de instrumentos que permitirán el análisis descriptivo de una variable. En primer lugar se indicará la forma de organizar y presentar la información, una vez que se ha observado la población y ha sido medido uno de los caracteres de todos y cada uno de los elementos de la misma. Esta operación nos llevará a la obtención de una distribución de frecuencias. Una vez que se tienen los datos organizados mediante esa distribución hay que iniciar el proceso de análisis de la variable. Para ello, el primer instrumento al que se puede recurrir, tanto por su sencillez como por lo fácil de su interpretación, es la representación gráfica de ese carácter. Con la representación gráfica de la variable o del atributo, según proceda en cada caso, se consigue tener una visión de conjunto del fenómeno estudiado de una forma más rápida y perceptible que con la sola inspección numérica de la tabla o distribución. Para continuar este proceso de análisis de una variable hay que definir ciertos instrumentos que nos permitan estudiar sus características más relevantes. Entre las mismas cabe destacar las siguientes: 1) forma de la distribución; 2) medidas de posición (valor central o promedios); 3) dispersión; 4) asimetría; 5) curtosis.

2.2 Distribuciones de frecuencias unidimensionales.

El adjetivo unidimensional hace referencia a que en nuestro análisis solo se tendrá en cuenta un carácter. Al mismo se le va a representar simbólicamente mediante la letra X, mientras que para sus posibles realizaciones (valores o modalidades, según se trate de variable o de un atributo, respectivamente) se utilizará la letra x minúscula.

En lo sucesivo se entenderá que el carácter observado es de tipo cuantitativo y que, en consecuencia, estamos trabajando con variables. En realidad el tratamiento que se le da a los atributos, en cuanto a distribución de frecuencias, es muy similar al de las variables discretas.

Por *distribución de frecuencias se va a entender al conjunto de valores que ha tomado una variable con sus frecuencias correspondientes*. Simbólicamente, una distribución de frecuencias vendría dada por los pares (x_i, n_i), donde x_i son los valores de la variable y n_i son sus frecuencias. Hay que señalar, en esta definición, que la frecuencia asociada a un valor de la variable es el número de veces que se repite ese valor. A la misma se le conoce como frecuencia absoluta.

2.2.1 Distribuciones de frecuencias para valores no agrupados.

Pueden considerarse básicamente dos tipos de distribuciones de frecuencias. Aquellas en las que los valores de la variable no están agrupados y las que presentan esos valores agrupados en intervalos.

Las primeras se corresponden a variables discretas. Este tipo de variables se caracteriza, como ya se indicó en su momento, por tener un número finito de valores o infinito numerable, de forma que entre dos valores consecutivos no existe otro. Pero aunque estos valores sean observables resulta que, a veces, el número de ellos es tan elevado que resulta aconsejable presentar la distribución o tabla estadística con los valores de la variable agrupados en intervalos. Esta forma de proceder podría llevarnos a pensar que estamos trabajando con variables continuas, cuando en realidad no lo son, pues en el caso de éstas, a diferencia de las primeras, dentro de cualquier intervalo de valores se pueden considerar que hay infinitos valores distintos.

La forma estándar de dar una distribución de frecuencias con valores no agrupados es la que aparece en la Tabla 1. Las frecuencias de esta tabla puede ser unitarias o mayores que uno. El primer caso tienen poco interés para la Estadística, pues como ya se indicó en la capitulo primero, el objeto de la misma era el estudio de colectivos grandes y nunca las poblaciones con un número de elementos muy reducido.

Junto a las frecuencias absolutas de los valores de una variable resulta habitual dar, también, lo que se conoce como frecuencias relativas. Para un valor concreto, la frecuencia relativa, que representaremos por f_i, es el cociente entre la frecuencia absoluta y el numero total de observaciones N. Es decir, $f_i = n/N$. Estas frecuencias se puede expresar en porcentajes o en tantos por uno. A su vez, las frecuencias, tanto las absolutas

como las relativas, se puede dar de forma acumulada. Las frecuencias absolutas acumuladas se representan por N_i y las relativas acumuladas por F_i.

Tabla 1. Distribución de frecuencias para valores no agrupados.

Valores de la variable x_i	Frecuencias Absolutas n_i	Frecuencias Relativas $f_i = n_i / N$.	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas $F_i = N_i / N$
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_i	n_i	f_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_k	n_k	f_k	$N_k = N$	$F_k = 1$
	$\sum_i n_i = N$	$\sum_i f_i = 1$		

***Ejemplo 1**. En la tabla adjunta se da la distribución de las 500 hogares de un barrio según el número de sus miembros.*

Tamaño de los hogares xi	*Nº de hogares ni*	f_i	N_i	F_i
1	*40*	*0,08*	*40*	*0,08*
2	*70*	*0,14*	*110*	*0,22*
3	*110*	*0,22*	*220*	*0,44*
4	*90*	*0,18*	*310*	*0,62*
5	*48*	*0,096*	*358*	*0,716*
6	*42*	*0,084*	*400*	*0,8*
7	*40*	*0,08*	*440*	*0,88*
8	*35*	*0,07*	*475*	*0,95*

9	*20*	*0,04*	*495*	*0,99*
10	*5*	*0,01*	*500*	*1*
	500	*1*		

Esta distribución, además de dar los valores de la variable y sus frecuencias absolutas, recoge las relativas y las acumuladas. Se trata de la distribución de una variable discreta con un número finito de valores.

2.2.2 Distribuciones de frecuencias para valores agrupados.

Este tipo de distribuciones se asocia, fundamentalmente, con variables continuas, aunque, como ya se ha señalado con anterioridad, en algunos casos también es aplicable a variables discretas, especialmente en aquellas situaciones en las que la variable toma muchos valores, de forma que si éstos nos e agruparan, la tabla resultaría demasiado extensa y la función de síntesis de la misma se perdería.

La elaboración de la distribución de frecuencias de una variable continua plantea algunos problemas que no se dan en el caso de variables discretas. Se trata de decidir el número de intervalos en los que hay que agrupar los valores de la variable así como si la amplitud o recorrido de los mismos debe ser igual. Estas cuestiones no tienen una respuesta determinada de antemano. La solución dependerá de cada caso concreto, por lo que no tiene sentido entrar en la casuística de las distintas situaciones que pudieran darse. Otro problema surge cuando un valor de la variable coincide exactamente con un extremo del intervalo, con lo que hay dudas sobre su inclusión en ese intervalo o el siguiente. Como solución a este problema es habitual proceder a definir intervalos abiertos por la izquierda y cerrados por la derecha, lo que implica que intervalo definido entre *a* y *b* incluye a todos los valores menores o iguales que *b* pero mayores que *a*.

En general, una distribución de frecuencias para una variable continua será como la que se da en la Tabla 2. A partir del contenido de esta tabla hay que definir los siguientes conceptos:

a) Amplitud del intervalo. Es la diferencia entre el extremo superior y el inferior. Así para el intervalo *i-ésimo*, la amplitud vendría dada por:

$$a_i = L_i - L_{i-1} \qquad (2.1)$$

b) Marca de clase. Es el punto central de cada intervalo. Esta se suele representar por x_i. Para el intervalo *i-ésimo* viene será:

$$x_i = (L_i + L_{i-1})/2 \qquad (2.2)$$

Tabla 2. Distribución de frecuencias para valores agrupados

Variable (intervalos $L_{i-1} - L_i$)	Amplitud a_i	Marca de clase x_i	Frec. abs. n_i	Frecuencia relativa f_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa Acum. F_i
$L_0 - L_1$	a_1	x_1	n_1	n_1/N	$N_1= n_1$	N_1/N
$L_1 - L_2$	a_2	x_2	n_2	n_2/N	$N_2= n_1 + n_2$	N_2/N
$L_2 - L_3$	a_3	x_3	n_3	n_3/N	$N_3= n_1+ n_2 +n_3$	N_3/N
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
$L_{i-1} - L_i$	a_i	x_i	n_i	n_i /N	$N_i =n_1+ n_2+...+n_i$	N_i / N
.	.	.	.	.	.	.
.	.	.	.	.	.	.
$L_{k-1} - L_k$	a_k	x_k	n_k	n_k /N	$N_k =n_1+ n_2+...+n_k = N$	$N_k /N =1$
			$\sum_i n_i = N$	$\sum_i fi =1$		

Esta agrupación de los valores de la variable en intervalos, aunque resulta operativamente necesaria, conlleva un problema grave que se conoce como error de agrupamiento. Este error es la consecuencia directa de la pérdida de información provocada al incluir un conjunto de observaciones en un mismo intervalo. Así, antes de agrupar las n_i observaciones en el intervalo *i-ésimo*, se sabe cuales son los valores concretos observados y sus respectivas frecuencias individuales. Ahora bien, cuando esos valores se agrupan en un intervalo se pierde esa información individualizada. En esos casos lo que se hace es sustituir a todos y cada uno de ellos por su valor medio que viene representado por la marca de clase. Pero esta solución, como se verá más adelante, implica asumir ciertos supuestos que nos inducen a error. Este error es el coste de la pérdida de información que se causa por el agrupamiento de las valores de la variable.

Ejemplo 2 *La distribución del presupuesto semanal en alimentación de un conjunto de 265 familias expresado en euros es el que figura en la tabla siguiente:*

Presupuestos L_{i-1} - L_i	a_i	x_i	Familias n_i	f_i	N_i	F_i
80-100	20	90	10	0,0377	10	0,0377
100-110	10	105	35	0,1321	45	0,1698
110-115	5	112,5	40	0,1509	85	0,3208
115-120	5	117,5	45	0,1698	130	0,4906
120-130	10	125	55	0,2075	185	0,6981
130-150	20	140	30	0,1132	215	0,8113
150-170	20	160	20	0,0755	235	0,8868
170-210	40	190	15	0,0566	250	0,9434
210-270	60	240	10	0,0377	260	0,9811
270-360	90	315	5	0,0189	265	1,0000
Total			**265**	**1**		

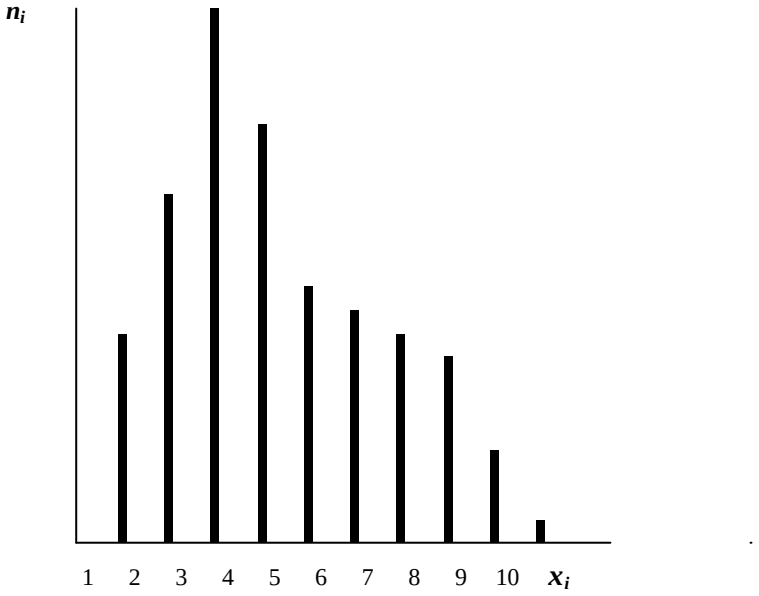
En este caso se trata de una variable continua con sus valores agrupados en intervalos cuya amplitud es variable. Este tipo de intervalos permite tratar de forma distinta a los valores de la variable, según donde se localice la mayor parte de las observaciones. En este sentido la amplitud de los intervalos es inversa a la frecuencia de los mismos. Esta forma de proceder evita que la mayor parte de las observaciones se concentre en un solo intervalo o en unos pocos.

2.3 Análisis gráfico

El tipo de representación gráfica depende en gran medida de la naturaleza del carácter de los elementos de la población con el que se esté trabajando. Así, si se trata de una variable se recurrirá al diagrama de barras en el caso de que sea discreta y sus valores no estén agrupados. Este diagrama se realiza haciendo uso de un sistema cartesiano en el que sobre el eje de abscisas se ponen los valores de la variable y sobre el de ordenadas las frecuencias, tanto absolutas (n_i) como relativas (f_i). Un ejemplo de este tipo de gráfico es el que se da en la Figura 1, donde se han representado los datos del Ejemplo 1. Hay que señalar que la anchura de las barras será, en cualquier caso, una cuestión de estética, pues la misma no hace alusión, en ningún caso, ni al valor de la variable ni al frecuencia del mismo. Dicho de otra forma, la superficie de la misma es independiente de la magnitud que se representa. En este sentido tan válido es el diagrama dado en la Figura 1 como el de la Figura 2.

Tanto en la Figura 1 como en la Figura 2 se ha representado las frecuencias absolutas. Pero estas figuras no habrían cambiado para nada si en su lugar si se hubiera trabajado con las frecuencias relativas. La única diferencia es que el eje de ordenadas tomaría como valor máximo la unidad, pero la proporción entre las barras no cambia de un gráfico a otro.

Figura 1. Diagrama de barras para la distribución del Ejemplo 1.



Otra represtación gráfica que también puede realizarse con los datos de una variable discreta es lo que se conoce como un diagrama escalonado o acumulativo. En este caso, sobre el eje de abscisas se siguen llevando los valores de la variable, mientras que sobre el de ordenadas se colocan las frecuencias acumuladas, bien absolutas (N_i) o relativas (F_i). En la figura 3 se ha representado el diagrama escalonado para la variable del Ejemplo 1.

Figura 2. Diagrama de barras para los datos del Ejemplo 1.

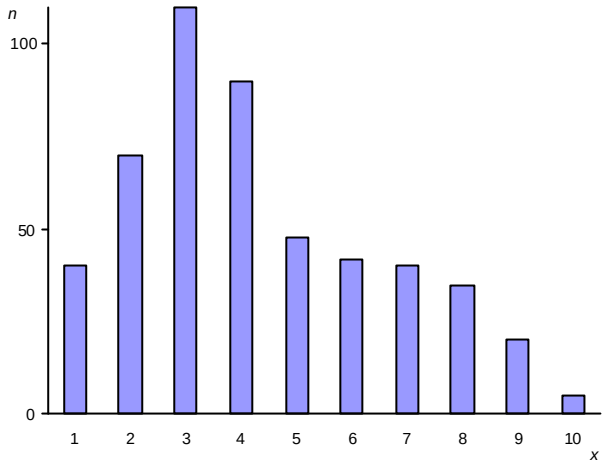
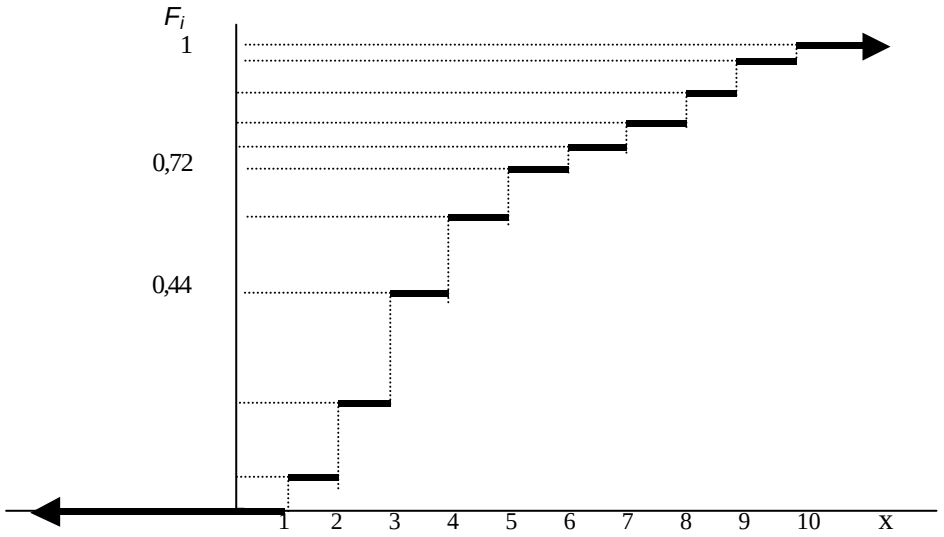


Figura 3. Diagrama escalonado correspondiente a los datos del Ejemplo 1.



Frente a este tipo de gráfico, cuando la naturaleza de la variable sea continua, entonces la representación gráfica más adecuada es el histograma o también conocido como histograma de frecuencias. Este tipo de gráficos podría utilizarse también en los casos de variables discretas con valores agrupados, aunque no resulta aconsejable hacer uso de

los histogramas para variables discretas por los problemas que conlleva asimilar una variable discreta a otra de tipo continuo.

Un histograma se realiza también haciendo uso de un sistema cartesiano, donde sobre el eje de abscisas se llevan los valores de la variable. Pero ahora ya no se trata de valores puntuales, sino de intervalos, y sobre éstos se levantan rectángulos, que tienen por base la amplitud del intervalo y por altura su frecuencia. El área de esos rectángulos deberá ser siempre proporcional a la frecuencia, de manera que cuando la amplitud de los intervalos no sea constante, entonces la altura de los rectángulos no será la frecuencia sino lo que se conoce como densidad de frecuencia definida de la forma siguiente:

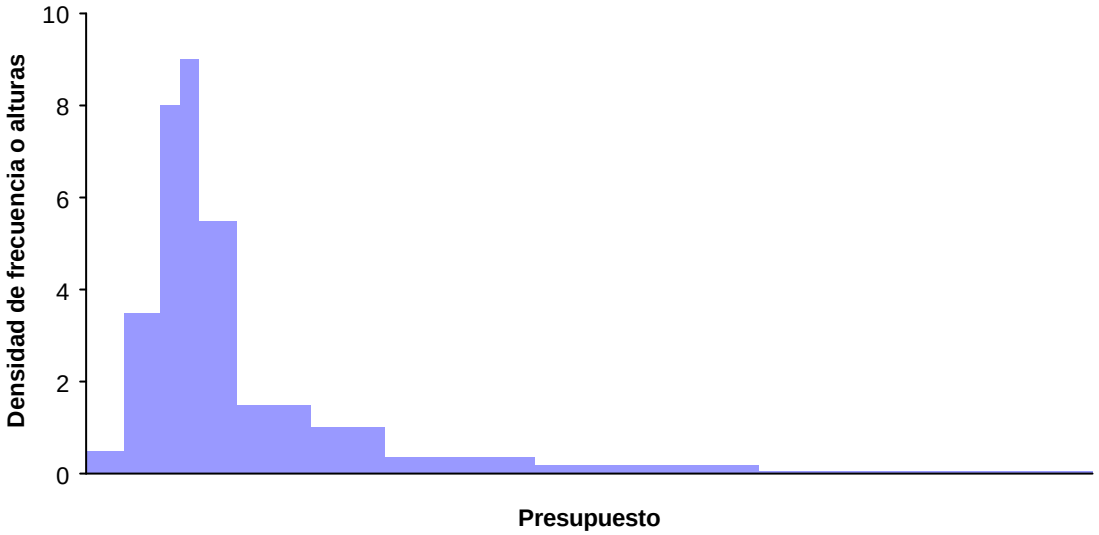
$$h_i = \frac{n_i}{a_i} \qquad i = 1,2,3,....,k \qquad (2.3)$$

La Figura 4 recoge el histograma correspondiente a los datos del Ejemplo 2. En este caso se ha procedido a calcular las correspondientes densidades de frecuencias también conocidas como alturas, dado que los intervalos son de amplitud variable. Los datos numéricos que se han representado son los que aparecen en la Tabla 3.

Tabla 3. Distribución de los presupuestos familiares.

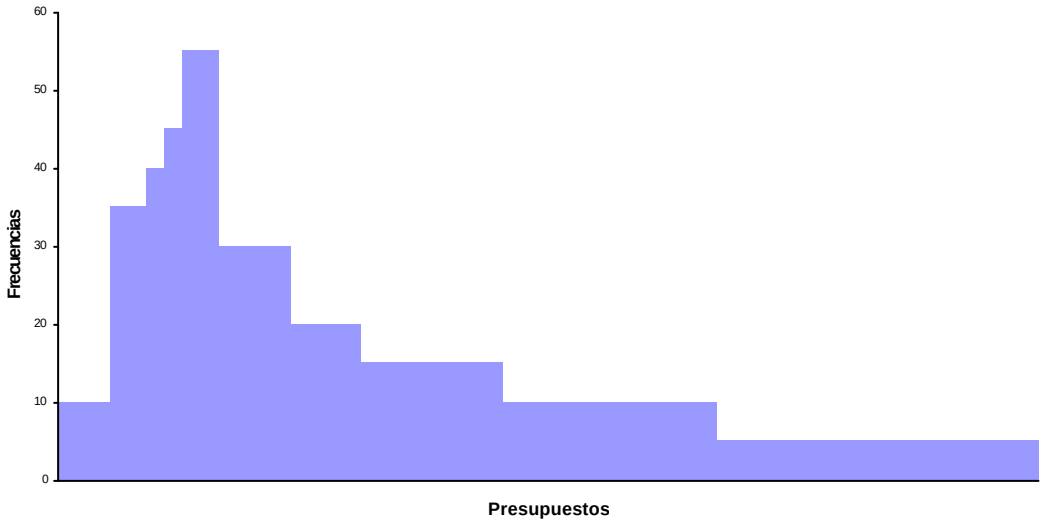
Presupuestos L_{i-1} - L_i	a_i	Familias n_i	h_i
80-100	20	10	0,50
100-110	10	35	3,50
110-115	5	40	8,00
115-120	5	45	9,00
120-130	10	55	5,50
130-150	20	30	1,50
150-170	20	20	1,00
170-210	40	15	0,38
210-270	60	10	0,17
270-360	90	5	0,06
Total		**265**	

Figura 4. Histograma para los datos del Ejemplo 2.



Si en lugar de trabajar con las alturas se hubieran llevado sobre el eje de ordenadas directamente las frecuencias, entonces el histograma correspondiente es el que aparece en la Figura 5. Como puede apreciarse, éste es muy distinto del anterior. Este último no es correcto porque el área de cada rectángulo no es proporcional a las frecuencias y, en consecuencia, muestra una realidad distorsionada.

Figura 5. Histograma para los datos del Ejemeplo 2.

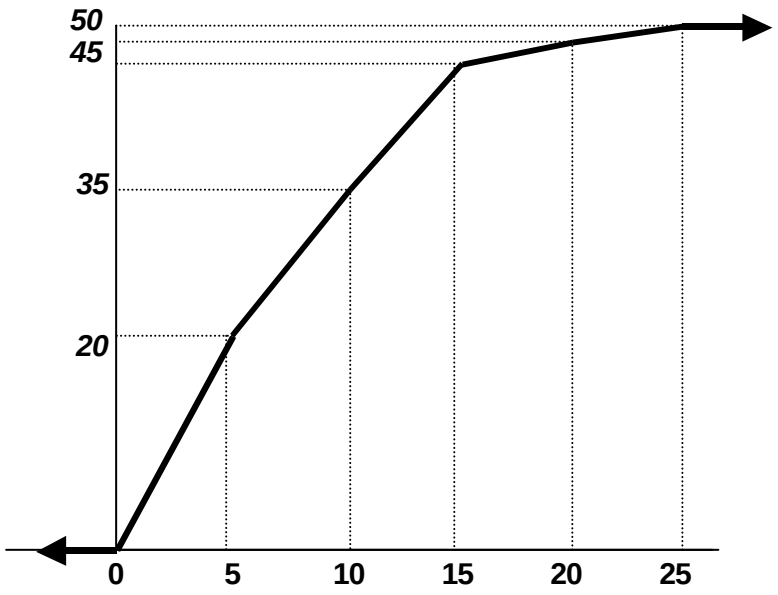


Al igual que para variables discretas se definió el diagrama escalonado para representar las frecuencias acumuladas, para el caso de continuas se puede proceder de forma similar. Pero en este caso, a la gráfica correspondiente, se le conoce como polígono de frecuencias acumuladas.

***Ejemplo 3**. A 50 clientes de una institución financiera se les ha preguntado por el tiempo que han tenido que esperar en la cola de la caja para realizar una gestión. Sus respuestas se han organizado en la siguiente tabla.*

Tiempo	*Clientes*	N_i
0-5	*20*	*20*
5-10	*15*	*35*
10-15	*10*	*45*
15-20	*3*	*48*
20-25	*2*	*50*

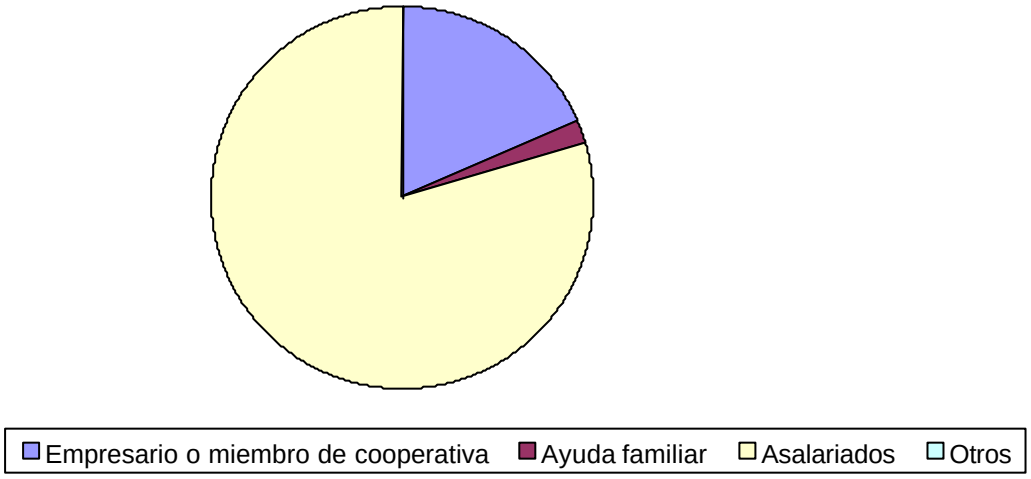
Obtenga el polígono de frecuencias acumuladas.



Una vez que se han señalado los instrumentos gráficos más habituales en el análisis de variables, quedan por introducir los correspondientes cuando de lo que se trata es de atributos. Ahora, las modalidades no tienen la dimensión numérica de los valores de las variables. Esto hace que haya que pensar en otros tipo de gráficos. Entre los más

utilizados están los diagramas de tarta. En estos casos lo que se hace es asignarle a cada modalidad del atributo un sector circular proporcional a su frecuencia. Para aclarar esta idea en la Figura 6 se ha representado la distribución de la población ocupada en España según su situación profesional, considerando cuatro modalidades distintas.

Figura 6. Población ocupada en España según situación profesional en 2000. (Miles de personas)

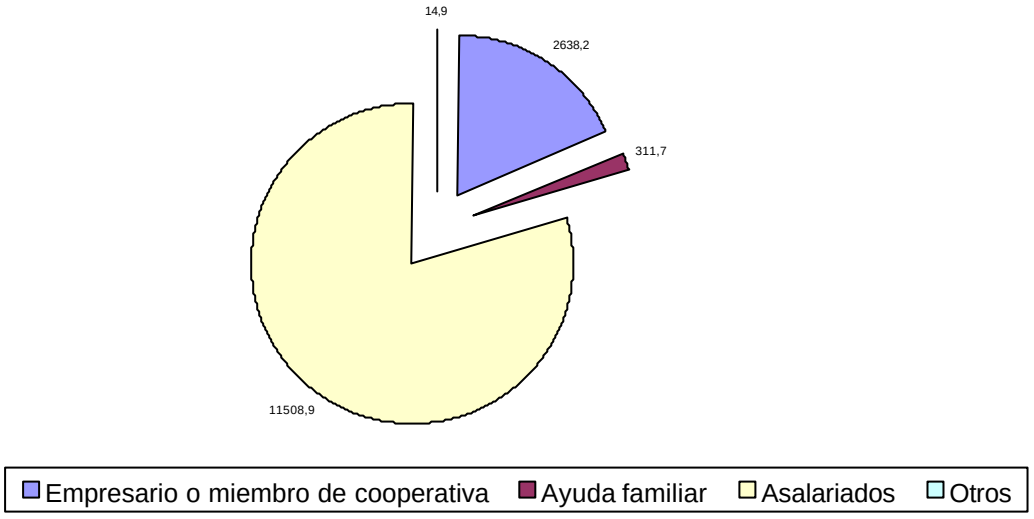


Fuente: EPA. Página web del INE

Este tipo de gráficos permite visualizar de forma bien clara cual es la estructura de un atributo concreto como es en este caso la situación profesional de la población ocupada, donde, como puede apreciarse la mayor parte de los ocupados son asalariados. Además estos gráficos se prestan a que puedan realizarse comparaciones para el mismo atributo en distintos ámbitos espaciales (por ejemplo, España y la situación en las CC.A.A. o con respecto a otros países) o temporales (por ejemplo, el año 2000 con respecto a años anteriores).

Cuando, como en el caso anterior, una de las modalidades no se aprecia porque su frecuencia es muy pequeña, entonces el gráfico se puede presentar como se hace en la Figura 7.

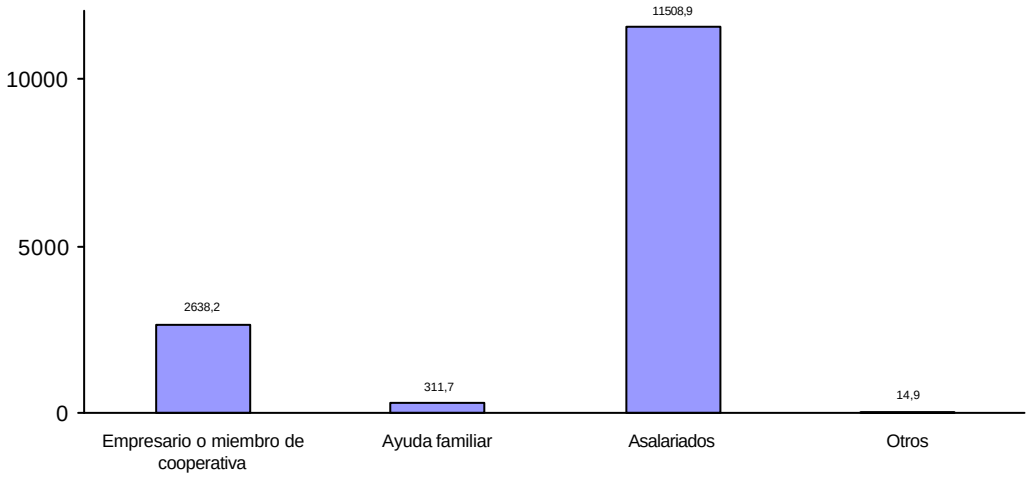
Figura 7. Población ocupada en España según situación profesional en 2000. (Miles de personas)



Fuente: EPA. Página web del INE

Este mismo atributo se ha representado en la Figura 8 haciendo uso de un diagrama de rectángulos. Este instrumento gráfico es muy similar al diagrama de barras visto para variables discretas.

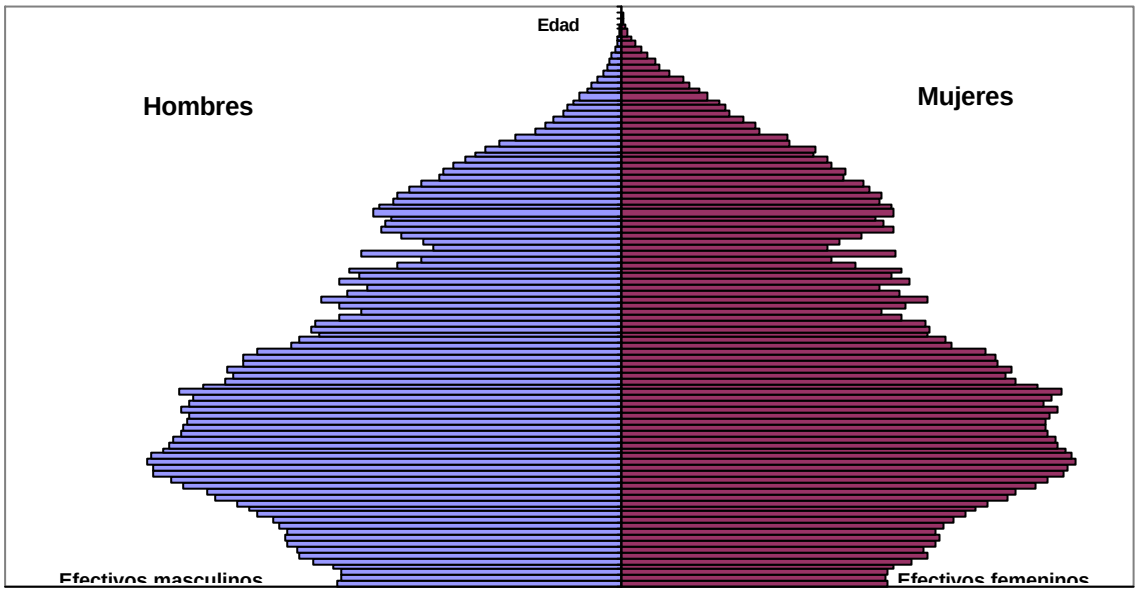
Figura 8. Población ocupada en España según situación profesional en 2000. (Miles de personas)



Fuente: EPA. Página web del INE

Pero este repertorio de gráficos no agota las posibilidades de representación. A los mismos se le puede añadir los pictogramas, cartogramas, etc. Sin embargo, los señalados son los que se utilizan con más frecuencia y, en cualquier caso, según el atributo con el que se esté trabajando habrá que seleccionar el más adecuado de entre la amplia gama de tipos de gráficos existentes.

Figura 9 . Pirámide de la población de Andalucía. 1998



Para finalizar este apartado dedicado a las representaciones gráficas vamos a dedicar unas líneas a un gráfico que tiene la particularidad de que en el mismo se hace uso de una variable continua, como es la edad, y un atributo, como es el sexo de la población. Nos estamos refiriendo a las pirámides de población, instrumento gráfico de gran utilidad en Demografía. Se trata de dos histogramas que comparten el mismo eje de abcisas, sobre el cual se lleva la edad de la población. Por otro lado, sobre el eje de ordenadas se llevan los efectivos poblacionales, tanto de hombres como de mujeres. Pero se trata de efectivos expresados no en cifras absolutas sino en porcentajes o en cualquier potencia de diez. Una muestra de este tipo de gráficos es el recogido en la Figura 9, donde se muestra la estructura por sexo y edad de la población de Andalucía para 1998.

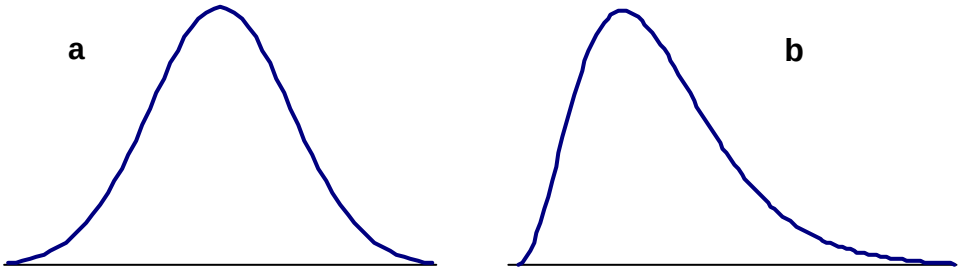
2.4 Forma de la distribución

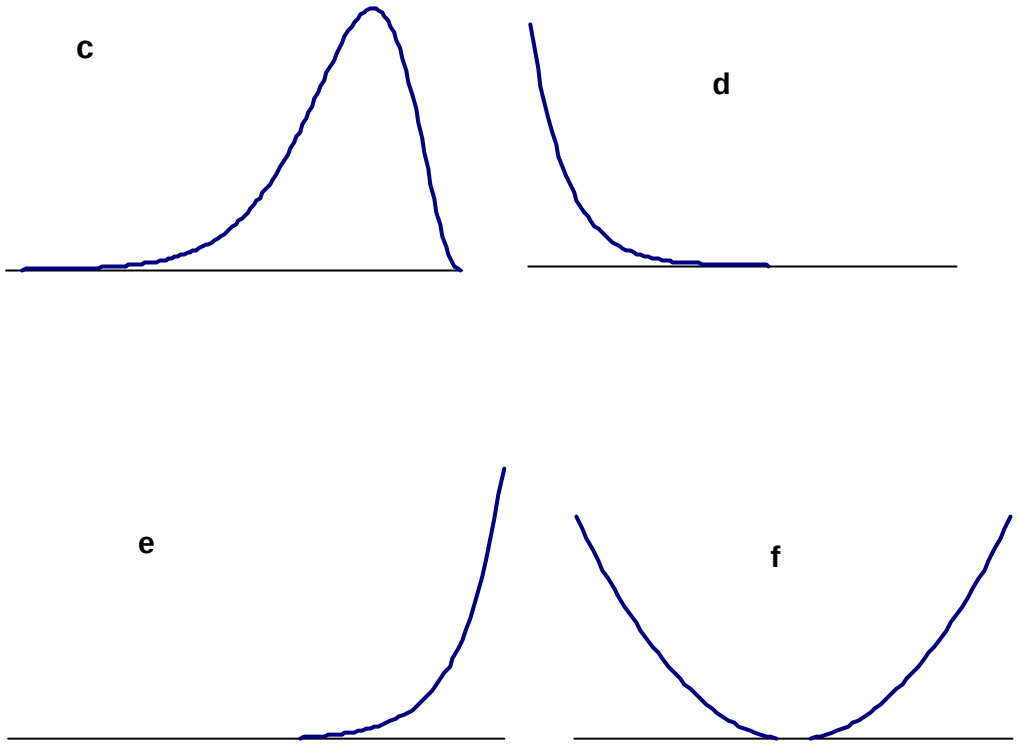
Para poder estudiar la forma de una distribución es preciso disponer de un número de observaciones de la variable suficientemente grande, como para poder determinar patrones de comportamiento o regularidades en dichas observaciones. Así, si se dispone de muy pocas observaciones, no tiene sentido hablar de la forma de la distribución.

Para estudiar la forma de la distribución la mejor herramienta de la que podemos hacer uso es de su representación gráfica, ya sea el diagrama de barras o el histograma. A partir de estas gráficas será posible, de forma fácil, concluir si las observaciones están concentradas en torno a unos pocos valores de la variable o si ocurre lo contrario, si la concentración tiene lugar en un extremo del recorrido de la variable o en el otro, etc.

Las formas más habituales de la distribución de una variable son las de tipo campaniforme (a), tanto simétricas como asimétricas (b, c). Otro formas también habituales son la que tienen forma de *jota*, invertida o normal (d, e), y las distribuciones en forma de U (f).

Figura 10. Formas de la distribución de una variable.





2.5 Promedios.

La reducción estadística que se consigue mediante la tabulación, en la mayoría de los casos, no resulta suficiente si lo que se persigue es que el "exceso" de información no nos impida ver lo que hay detrás de ella. Por tal razón, esa reducción hay que llevarla hasta el extremo de quedarnos con un solo dato que a su vez sea representativo de todo el conjunto. A ese único dato se le conoce de forma genérica como promedio. Con la obtención de promedios lo que se consigue es determinar cual es el nivel medio de la variable y, además, facilita las comparaciones entre variables. A los promedios se les conoce también como medidas de tendencia central. En principio, el único requisito que, de forma general, se le exige a cualquier promedio es que su valor esté comprendido entre los valores extremos de la variable. Con esta única condición, el número de promedios que pueden definirse es muy elevado, si bien los más habituales son la media aritmética, la media geométrica, la mediana, la moda, la media cuadrática, la media armónica, etc.

De todos ellos vamos a estudiar la media aritmética, la media geométrica, la mediana y la moda.

2.5.1 La media aritmética.

La media aritmética se define como aquel valor que resulta de dividir la suma de todas la observaciones entre el número de ellas. El resultado de este cociente no coincidirá necesariamente con valor alguno de la variable, pero si debe ser un valor del recorrido de la misma y vendrá expresado en las misma unidad de medida de esa variable. Por la forma en que se ha definido este promedio solo tiene sentido aplicarlo a variables de naturaleza cuantitativa, pues sus valores se pueden sumar, pero no las modalidades de un atributo.

Formalmente, si x_i es el valor *i-ésimo* de la variable X, siendo n_i el número de veces que se presenta ese valor, entonces la media se define como:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} \qquad (2.4)$$

Ejemplo 1: *Sea X una variable que representa el volumen de facturación de una empresa a lo largo de los 25 días laborales de un mes:*

x_i (miles €)	n_i
10,5	2
12,4	3
13,2	9
14,8	6
15,8	4
16,5	1

En este caso la media vendrá dada por:

$$\bar{x} = \frac{(10,5)(2)+(12,4)(3)+(13,2)(9)+(14,8)(6)+(15,8)(4)+(16,5)(1)}{25} = 13,82 \quad miles \quad €.$$

El ejemplo anterior es válido solo en el caso de distribuciones con datos no agrupados. Si los datos estuvieran agrupados, entonces los valores individuales de la variable son desconocidos, lo que impide hacer uso de la expresión cálculo anterior. Para dar solución a este problema se procede asumiendo la hipótesis de que todas las observaciones de un intervalo se distribuyen uniformemente dentro del mismo, por lo que es admisible sustituir todos los valores del intervalo por su marca de clase. Si se opta por esta solución, entonces es posible hacer uso de la expresión dada en *(2.4)* para el cálculo de la media aritmética.

Ejemplo 2: *Si la para la empresa anterior, en lugar de considerar las ventas en 25 días, consideramos las ventas en 300 días, entonces los datos serían los de la tabla siguiente:*

Ventas en pesetas (miles de €)	x_i	n_i
10,0 –12,0	*11,0*	*40*
12,0 – 13,0	*12,5*	*60*
13,0 – 14,0	*13,5*	*110*
14,0 – 15,0	*14,5*	*50*
15,0 – 16,0	*15,5*	*30*
16,0 – 18,0	*17,0*	*10*

Ahora la nueva media sería:

$$\bar{x} = \frac{(11)(40)+(12{,}5)(60)+(13{,}5)(110)+(14{,}5)(50)+(15{,}5)(30)+(17{,}0)(10)}{300} = 13{,}45 \quad miles \quad €$$

Esta forma de obtener la media aritmética implica que no se calcula de forma exacta, pues la frecuencia no tiene porque distribuirse de manera uniforme dentro del intervalo, por lo que al proceder de esa forma se comete un error que se le conoce como error de agrupamiento. Este viene condicionado por el número de intervalos que se estén considerando, así como por el tamaño de la población. Veamos este concepto de forma numérica.

Ejemplo 3. *Obtener la media de la variable del Ejemplo 1 pero con los datos agrupados según los intervalos de Ejemplo 2.*

En este caso las 25 observaciones se presentan en la tabla siguiente:

Ventas en pesetas (miles de €)	x_i	n_i
10,0 –12,0	*11,0*	*2*
12,0 – 13,0	*12,5*	*3*
13,0 – 14,0	*13,5*	*9*
14,0 – 15,0	*14,5*	*6*
15,0 – 16,0	*15,5*	*4*
16,0 – 18,0	*17,0*	*1*

Ahora, la media aritmética sería:

$$\bar{x} = \frac{(11)(2)+(12,5)(3)+(13,5)(9)+(14,5)(6)+(15,5)(4)+(17,0)(1)}{25} = 13,88 \quad miles \quad €$$

Como puede apreciarse el valor de la media ha cambiado, pasando de 13,82 a 13,88. La diferencia entre ambos es el error de agrupamiento que se ha cometido como consecuencia de trabajar con datos agrupados en intervalos.

Si lo que se persigue es obtener la media de una variable en la que los valores de la misma no tienen todos ellos la misma importancia o significación, entonces se procede a obtener la **media aritmética ponderada**, en la que cada valor de la esa variable se multiplica por sus respectivo peso o ponderación (w_i) que refleja la importancia de ese valor, pero que no es su frecuencia. Si la suma de esos productos la dividimos por la suma de las ponderaciones, lo que se obtiene es la media aritmética ponderada.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \qquad (2.5)$$

Ejemplo 4 *Un alumno ha realizado un examen que constaba de cinco preguntas. En cada una de ellas ha obtenido las siguientes puntuaciones: 5; 6,5; 7; 8 y 7,5. Obtenga la nota final del examen si las ponderaciones de esas preguntas son: 0,1; 0,25; 0,15; 0,25 y 0,25.*

$$\bar{x} = \frac{(5)(0{,}1)+(6{,}5)(0{,}25)+(7)(0{,}15)+(8)(0{,}25)+(7{,}5)(0{,}25)}{1} = 7{,}05$$

frente a una media aritmética simple de 6,8.

2.5.1.1 Propiedades de la media aritmética.

1ª La suma de las desviaciones de los valores de la variable respecto de la media vale cero. Es decir: $\sum_{i=1}^{k} \left(x_i - \bar{x}\right) n_i = 0$

La demostración de esta propiedad es como sigue:

$$\sum_{i=1}^{k} \left(x_i - \bar{x}\right) n_i = \sum_{i=1}^{k} x_i n_i - \bar{x} \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} x_i n_i - N\,\bar{x} = \sum_{i=1}^{k} x_i n_i - \sum_{i=1}^{k} x_i n_i = 0$$

2ª La media aritmética no varía si todas las frecuencias de su distribución se multiplican o dividen por una constante.

Llamemos *C* a constante por la que se van a multiplicar todas la frecuencias, de tal forma que $n_{Ci} = Cn_i$. En tal caso la media será:

$$\bar{x}_C = \frac{\sum_{i=1}^{k} x_i n_{Ci}}{N_C} = \frac{\sum_{i=1}^{k} x_i Cn_i}{CN} = \frac{C\sum_{i=1}^{k} x_i n_i}{CN} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \bar{x}$$

3ª La suma de las desviaciones al cuadrado de los valores de la variable respecto de una constante C cualquiera se hace mínima, $S(C) = \sum_{i=1}^{k} (x_i - C)^2$ *, cuando esa constante C coincide con la media aritmética (Teorema de Konig):*

Para demostrar esta propiedad basta con minimizar esa suma de desviaciones al cuadrado con respecto a *C*. El valor de esa constante *C* será el que anule la primera derivada y haga que la segunda sea positiva.

$$\frac{dS(C)}{dC} = \frac{d\left(\sum_{i=1}^{k} (x_i - C)^2\right)}{dC} = 2\sum_{i=1}^{k} (x_i - C)(-1) = 0$$

A partir de este resultado se obtiene que:

$$\sum_{i=1}^{k} x_i = NC$$

de forma que si se dividen ambos miembros de la igualdad por *N* queda demostrada la tercera propiedad de la media.

4ª Si a todos los valores de una variable le sumamos una constante C, la media aritmética queda aumentada en dicha constante. Es decir, la media se ve afecta por cambios de origen en la variable.

Para demostrar esta propiedad vamos a definir una nueva variable $y_i = x_i + C$. Ahora se obtendrá la media de y_i.

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum (x_i + C) n_i}{N} = \frac{\sum x_i n_i + NC}{N} = \bar{x} + C$$

5ª Si todos los valores de una variable los multiplicamos por una constante C, la media aritmética queda multiplicada por dicha constante. Es decir, la media se ve afecta por cambios de escala en la variable.

Al igual que en caso anterior, definamos una nueva variable $y_i = Cx_i$. Ahora se obtendrá la media de y_i.

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum Cx_i n_i}{N} = \frac{C\sum x_i n_i}{N} = C\bar{x}$$

6ª Si de un conjunto de valores obtenemos dos o más subconjuntos disjuntos, la media aritmética de todo el conjunto se relaciona con todas las medias aritméticas de los diferentes subconjunto disjuntos.

Supongamos que agrupamos las N observaciones en K subconjuntos disjuntos:

$$\left(x_{11}, x_{12}, ..., x_{1N_1}\right), \left(x_{21}, x_{22}, ..., x_{2N_2}\right), ..., \left(x_{K1}, x_{K2}, ..., x_{KN_K}\right)$$

Ahora la media es:

$$\bar{x} = \frac{\left(x_{11} + x_{12} + ... + x_{1N_1}\right) + \left(x_{21} + x_{22} + ... + x_{2N_2}\right) + ... + \left(x_{K1} + x_{K2} + ... + x_{KN_K}\right)}{N_1 + N_2 + ... + N_K} =$$

$$= \frac{\sum_{i=1}^{N_1} x_{1i} + ... + \sum_{i=1}^{N_K} x_{Ki}}{N_1 + N_2 + ... + N_K} = \frac{\bar{x}_1 N_1 + ... + \bar{x}_K N_K}{N_1 + N_2 + ... + N_K}$$

7ª La media es el centro de gravedad de la distribución.

Ventajas e inconvenientes de la media aritmética

Las principales ventajas son las siguientes:

1ª Hace uso de todos los valores para su cálculo

2ª Se puede calcular siempre.

3ª Es única.

Frente a estas ventajas el principal inconveniente es que se trata de un promedio sensible a los valores extremos de la variable, lo que en algunas ocasiones invalida su utilidad.

2.5.2 La media geométrica.

Sea una distribución de frecuencias (x_i, n_i). La media geométrica G se define como la raíz N-ésima del producto de los N valores de la distribución.

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \ldots . x_n^{n_n}} = \sqrt[N]{\prod_{i=1}^{N} x_i^{n_i}} \qquad (2.6)$$

La media geométrica goza de la propiedad de que su logaritmo es igual a la media aritmética de los logaritmos de los valores de la variable. También la media geométrica es siempre menor o igual que la media aritmética.

Ejemplo 5. *Las tasas de crecimiento de la economía de un país durante diez años son las que aparecen en la tabla siguiente:*

Tasas (X) en %	***Años n_i***
1	*2*
2	*2*
3	*3*
4	*2*
5	*1*

Obtener la tasa media anual de crecimiento.

Para esta variable, y dada su naturaleza, el promedio más adecuado es la media geométrica.

$$G = \sqrt[10]{\left((0{,}01)^2\,(0{,}02)^2\,(0{,}03)^3\,(0{,}04)^2\,(0{,}05)^1\right)} = 0{,}02475$$

Se trata de un promedio que, para su cálculo, al igual que la media aritmética, hace uso de toda la información de la variable. Sin embargo es menos sensible a los valores extremos de lo que lo es la media aritmética. Frente a estas ventajas o virtudes, este nuevo promedio tiene algunas limitaciones. Entre ellas destacaremos: a) es menos intuitivo que la media aritmética; b) su cálculo no es tan inmediato; c) en ocasiones no queda determinada. Si algún valor de la variable es nulo, entonces G se anula. Si la variable toma valores negativos este promedio da problemas.

La media geométrica se utiliza especialmente para promediar porcentajes, tasas, números índices, etc., y siempre que la variable presente variaciones acumulativas.

Ejemplo 6. *La población de una determinada provincia durante los años que se indica fue la siguiente:*

Año	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Población	*375,1*	*385,7*	*390,6*	*410,5*	*430,3*	*450,7*

Obténgase la tasa de crecimiento medio anual.

Para obtener la tasa pedida se podría proceder de forma distintas. Supongamos en primer lugar que calculamos inicialmente las tasas de crecimiento anuales. Estas tasas[1]*, que se definen como:* $\frac{P_t}{P_{t-1}} - 1$ *, son las que aparecen en la siguiente tabla:*

Año	*1996-95*	*1997-96*	*1998-97*	*1999-98*	*2000-99*
Tasa	*0,0282*	*0,0127*	*0,0509*	*0,0482*	*0,0474*

A partir de estas cinco tasas de crecimiento anual se podría calcular la media aritmética, pero este no sería el promedio más adecuado, dada la naturaleza de esta variable (se

[1] El concepto de tasa se expone de forma algo más amplia en el Capitulo 4 de este libro.

trata de tasas). En este caso el promedio más adecuado sería la media geométrica. Si a esa tasa media de crecimiento anual la llamamos r, entonces la misma vendría dada por:

$$r = \sqrt[5]{((0{,}0282)(0{,}0127)(0{,}0509)(0{,}0482)(0{,}0474))} = 0{,}03344$$

Otra forma de abordar el problema es la siguiente. Según se han definido las tasas de crecimiento se tiene que:

$$\frac{P_1}{P_0} = 1 + r_1$$
$$\frac{P_2}{P_1} = 1 + r_2$$
$$\frac{P_3}{P_2} = 1 + r_3$$
$$\frac{P_4}{P_3} = 1 + r_4$$
$$\frac{P_5}{P_4} = 1 + r_5$$

A partir de estas relaciones se llega a que $P_5 = P_0(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)$. *Ahora bien, si admitimos que durante todo el conjunto de años considerado la tasa de crecimiento anual ha sido la misma, entonces resulta que* $P_5 = P_0(1+r)^5$, *siendo r esa tasa de crecimiento medio anual constante. En estas condiciones, si igualamos los dos resultados tenemos que:* $(1+r)^5 = (1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)$, *por lo que finalmente se tiene que:*

$$r = \sqrt[5]{(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)} - 1$$

donde resulta que la tasa de crecimiento medio buscada se obtiene como un función de la media geométrica de las tasas de crecimiento anuales.

Este resultado se puede generalizar al caso de n periodos de tiempo, en cuyo caso se tendría:

$$r = \sqrt[n]{(1+r_1)(1+r_2)(1+r_3)(1+r_4).....(1+r_n)} - 1$$

Finalmente hay que indicar que si se procede de esta forma no es necesario calcular las tasas de crecimiento anual, pues r también se puede obtener de la forma siguiente:

$$r = \sqrt[n]{(1+r_1)(1+r_2)(1+r_3)(1+r_4).....(1+r_n)} - 1 = \sqrt[n]{\left(\frac{P_1}{P_0}\right)\left(\frac{P_2}{P_1}\right)\left(\frac{P_3}{P_2}\right)\left(\frac{P_4}{P_3}\right).....\left(\frac{P_n}{P_{n-1}}\right)} - 1 =$$

$$= \sqrt[n]{\frac{P_n}{P_0}} - 1$$

2.5.3 La mediana.

La mediana se puede definir de las siguientes formas:

1º *Es aquel valor de la distribución que ocupa el lugar central una vez los valores han sido ordenados de forma creciente, de menor a mayor*.

2º *Aquel valor de la distribución, una vez ordenada de menor a mayor, que deja a su izquierda y a su derecha el mismo número de observaciones*.

3º *El primer valor de la variable, una vez ordenados de menor a mayor, cuya frecuencia acumulada es mayor o igual que N/2*.

Con estas definiciones de la mediana *Me* lo que venimos a decir es que este promedio no se puede expresar mediante un fórmula.

Para determinar el valor de la *Me* de una distribución hay que proceder de forma distinta según se trate de distribuciones de frecuencias para variables discretas o continuas, pues en el primer caso los valores de la variable no están agrupados y sus frecuencias pueden ser unitarias o mayores que la unidad, mientras que para el segundo caso los valores se presentan siempre agrupados.

Cuando se trabaja con variables discretas con frecuencias unitarias, la mediana es el valor central una vez ordenados de menor a mayor. Si el número de observaciones fuera par, entonces la mediana sería la semisuma de los dos valores centrales.

Ejemplo 7. *Las notas de un examen de los diez alumnos de una clase son las siguientes: 3, 7, 6, 4, 8, 9, 5, 7, 4, 8. Obtener la nota mediana.*

En este caso lo primero que habría que realizar es ordenar de menor a mayor esos valores y como el número de ellos es par, entonces la mediana será la media de los valores centrales. En este caso se tendrá que:

$$Me = \frac{6+7}{2} = 6{,}5$$

Si se trata de una distribución para variables discretas con frecuencias mayores que la unidad y valores no agrupados, entonces el valor de la variable que corresponde a la primera frecuencia acumulada mayor o igual que N/2 será la mediana.

Ejemplo 8. *Las notas de estadística par los cien alumnos de primero de una facultad de económicas y empresariales son los que se dan a continuación:*

Notas (x_i)	***Alumnos (n_i)***	***N_i***
0	*1*	*1*
1	*3*	*4*
2	*5*	*9*
3	*7*	*16*
4	*10*	*26*
5	*28*	*54*
6	*19*	*73*
7	*12*	*85*
8	*8*	*93*
9	*5*	*98*
10	*2*	*100*

Obtener la nota mediana.

En este caso, dado que los valores de la variable no están agrupados en intervalos, la mediana se corresponde con el primer valor de la variable cuya frecuencia acumulada es mayor o igual que N/2. La forma práctica de obtener ese valor consiste en añadir a la

tabla original una columna adicional que recoja las frecuencias acumuladas. En este caso como N/2=50, entonces la nota mediana resulta ser 5.

Cuando se trabaja con datos agrupados, entonces los procedimientos anteriores no sirven. En estos casos se busca el intervalo mediano. El mismo será el primero cuya frecuencia acumulada sea mayor o igual que *N/2*. El valor de la mediana será uno de los valores comprendidos dentro de ese intervalo. Si admitimos que la frecuencia de ese intervalo se distribuye uniformemente dentro del mismo, entonces la mediana será aquel valor que sumado al extremo inferior del intervalo acumule un frecuencia igual a *N/2*. De acuerdo con este criterio, la mediana se podrá calcular, de forma aproximada, mediante la expresión (2.7).

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i \qquad (2.7)$$

Ejemplo 9. *Obtener la mediana para los datos del Ejemplo 8 si estos se dieran agrupados de la forma siguiente:*

Notas (x_i)	***Alumnos (n_i)***	***N_i***
0-2	*9*	*9*
2-4	*14*	*23*
4-6	***50***	***73***
6-8	*20*	*93*
8-10	*7*	*100*

Para determinar el valor de la mediana en este caso habría que proceder a su cálculo aproximado mediante la expresión dada en (2.7), para lo cual hay que determinar el primer lugar el intervalo mediano, que en este ejemplo es el que se corresponde con los valores 4-6 de la variable.

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 4 + \frac{50-23}{50} 2 = 5{,}08$$

Como puede observarse este valor no coincide con el obtenido previamente. Ello se debe al error de agrupamiento *motivado por la falta de información que conlleva no tener los datos desagregados, lo que nos impide conocer la frecuencia de cada uno de los valores de la variable.*

2.5.3.1 Propiedades de la *Me*

1ª *La mediana hace mínima la suma de todas las desviaciones absolutas*.

$$Min \sum_i |x_i - k| n_i = \sum_i |x_i - Me| n_i$$

2ª *La mediana se ve afecta por cambios de origen y cambios de escala*.

3ª *La mediana no se ve afectada si todas las frecuencias se multiplican o dividen por una misma constante*.

4ª La mediana no se influida por los valores extremos de la variable.

5ª *Para el caso de distribuciones campaniformes fuertemente asimétricas, la mediana resulta un promedio mejor que la media aritmética*.

6ª *Dado que en su cálculo no intervienen los valores extremos hace que se pueda obtener fácilmente incluso en presencia de intervalos abiertos* .

El principal inconveniente de la mediana es que, para su cálculo, no hace uso de toda la información que suministra la variable.

2 .5.4 La moda

La moda es aquel valor de la distribución que más se repite o que presenta mayor frecuencia.

De esta definición se deduce fácilmente que este promedio carece de interés en distribuciones de frecuencias unitarias, pues en esos casos todos los valores tienen idéntica frecuencia por lo que el valor modal no existe.

Este promedio se representa por *Mo.* Al igual que la mediana, tampoco tiene formula del cálculo.

Para su determinación se procede de forma distinta según se trate de distribuciones de frecuencias no agrupadas o agrupadas.

En el primer caso se aplica simplemente la definición. Así, en términos del *Ejemplo 8* tendremos que *Mo =5*, pues ese valor de la variable es el que presenta una frecuencia mayor. En este caso la *Me* y la *Mo* coinciden.

En el segundo caso se procede de la siguiente forma. Si todos los intervalos tienen la misma amplitud, entonces se busca el de mayor frecuencia (intervalo modal) y la moda será uno de los valores contenidos en el mismo. La forma aproximada de determinar ese valor, suponiendo de nuevo equidistribución de la frecuencia dentro del intervalo, será la siguiente:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i \qquad (2.8)$$

Ejemplo 10. *Obtener la moda para los datos del Ejemplo 9:*

Notas (x_i)	***Alumnos (n_i)***
0-2	*9*
2-4	*14*
4-6	***50***
6-8	*20*
8-10	*7*

Para determinar el valor de la moda en este caso habría que proceder a su cálculo aproximado mediante la expresión dada más arriba:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i = 4 + \frac{20}{14 + 20} 2 = 5{,}176$$

Como puede observarse, y al igual que ocurría en el caso de la mediana, este valor no coincide con el obtenido previamente. Ello se debe al error de agrupamiento *motivado por la falta de información que conlleva no tener los datos desagregados, lo que nos impide conocer la frecuencia de cada uno de los valores de la variable.*

Si la amplitud de los intervalos es distinta, entonces, en lugar de buscar el intervalo de mayor frecuencia, se busca el intervalo de mayor altura y se procede de igual forma que en la situación anterior:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i \qquad (2.9)$$

Ejemplo 11. *Obtener el salario mensual más frecuente, expresado en euros, para la siguiente distribución:*

Salarios (x_i)	***Asalariados (n_i)***	***a_i***	***$h_i = n_i / a_i$***
Menos de 500	*50*	*500*	*0,1*
De 500 a 900	*70*	*400*	*0,175*
De 900 a 1200	***120***	***300***	***0,4***
De 1200 a 1800	*100*	*600*	*0,1666*
De 1800 a 2700	*50*	*900*	*0,0555*
De 2700 a 5000	*20*	*2300*	*0,0087*

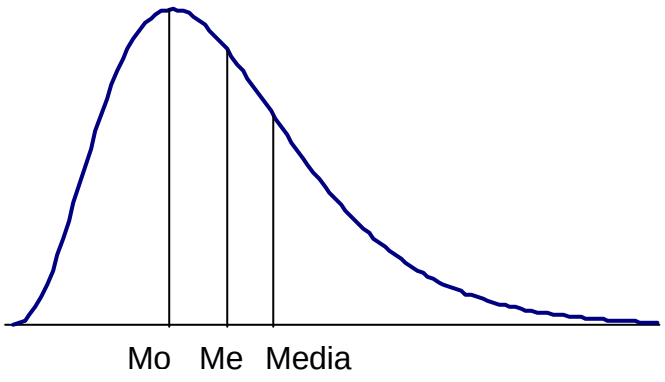
En este caso la moda será:

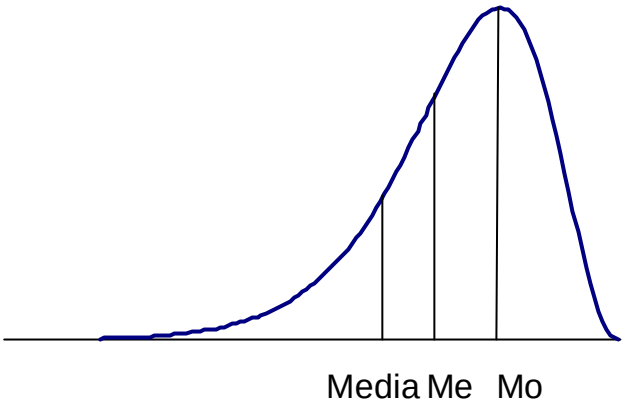
$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 900 + \frac{0{,}1666}{0{,}175 + 0{,}1666} 300 = 1046{,}34 \text{ €}$$

Este promedio pierde interés cuando la distribución tiene más de un máximo, es decir, cuando la distribución es bimodal o multimodal, pues, como ya se ha señalado con anterioridad, la finalidad de los promedios es resumir toda la información de la distribución en un solo dato. La moda tampoco tiene gran interés cuando la distribución no es campaniforme.

En distribuciones campaniformes simétricas se cumple que la media, la mediana y la moda toman el mismo valor ($\overline{x} = Mo = Me$). Pero si no son simétricas entonces los valores son distintos, y la relación entre los mismos es la que aparece en la Figura 11.

Figura 11. Relación entre la media aritmética, la mediana y la moda en distribuciones campaniformes asimétricas.





Las propiedades de la moda son muy similares a las de la mediana, en el sentido de que se ve afectada por cambios de origen y de escala, no cambia cuando las frecuencias se multiplican o dividen por una constante y no se ve afectada por los valores extremos de la distribución.

Como inconveniente hay que señalar que no hace uso de toda la información de la tabla y que en distribuciones multimodales pierde sentido.

2.6 Las cuantilas.

Previamente se ha definido la mediana como aquel valor de la variable que divide a la distribución en dos partes iguales. Siguiendo con esta idea podríamos plantearnos buscar valores de la variable que dividan la distribución en un determinado número de partes iguales. Todos tendrían la propiedad de que entre ellos siempre queda el mismo número de observaciones. Esta forma de proceder nos lleva al concepto de cuantilas. Estas son valores de la distribución que la dividen en partes iguales, es decir, en intervalos con igual número de observaciones. A todas ellas se les conoce genéricamente como medidas de posición no centrales.

Las cuantilas de uso más frecuente son las cuartilas, las decilas y las centilas.

En consonancia con la idea de cuantilas que se ha dado previamente, las definiciones concretas de cada una de ellas serían las siguientes:

Cuartilas (Q_i): son los tres valores de la variable que dividen la distribución en cuatro partes iguales, es decir, en cuatro intervalos dentro de cada cual están incluidos la cuarta parte de los valores u observaciones de la variable.

Decilas (D_i): son los nueve valores de la variable que dividen la distribución en diez partes iguales, es decir, en diez intervalos dentro de cada cual están incluidos la décima parte de los valores u observaciones de la variable.

Centilas (C_i): son los noventa y nueve valores de la variable que dividen la distribución en cien partes iguales, es decir, en cien intervalos dentro de cada cual están incluidos la centésima parte de los valores u observaciones de la variable.

Para obtener los valores de estas medidas se procederá de forma distinta según que la distribución esté agregada o no. En el caso de distribuciones discretas con valores no agregados se tendrá que:

Q_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/4*, para *i=1,2,3*.

D_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/10*, para *i=1,2,3,...9*.

C_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/100*, para *i=1,2,3,...,98,99*.

Si la distribución está agrupada en intervalos entonces la expresión general para determinar estas medidas, que es similar a la de la mediana, viene dada por:

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i} a_i \qquad (2.10)$$

en donde:

para *k=4* y *r=1,2,3*, tendremos las cuartilas
para *k=10* y *r=1,2,3,.....,9*, tendremos las decilas
para *k=100* y *r=1,2,3,.....,99* tendremos la centilas.

Ejemplo 12. *Con los datos del Ejemplo 11 obtener: a) el salario mínimo del 25% de los asalariados que más ganan; b) el salario máximo del 40% de los que menos ganan; c) el porcentaje de asalariados con salarios comprendidos entre la centila 25 y la decila 9.*

Salarios (x_i)	***Asalariados (n_i)***	***N_i***
Menos de 500	*50*	*50*
De 500 a 900	*70*	*120*
De 900 a 1200	*120*	*240*
De 1200 a 1800	*100*	*340*
De 1800 a 2700	*50*	*390*
De 2700 a 5000	*20*	*410*

a) En este caso lo que nos piden es que se obtenga la centila 75 o la tercera cuartila, pues ambos valores son iguales. Para determinar ese valor hay que determinar previamente $\frac{r}{k}N = \frac{75}{100}410 = \frac{3}{4}410 = 307,5$. *La primera frecuencia acumulada mayor que este valor se corresponde con la del intervalo "De 1200 a 1800", por lo que el valor buscado es:*

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i}a_i = C_{75} = Q_3 = 1200 + \frac{307,5 - 240}{100}600 = 1605\,€$$

b) Ahora el valor pedido se corresponde con la centila 40 o la cuarta decila. Al igual que antes, hay que determinar previamente $\frac{r}{k}N = \frac{40}{100}410 = \frac{4}{10}410 = 164$. *La primera*

frecuencia acumulada mayor que 164 es la correspondiente a la del intervalo "De 900 a 1200", así que el valor pedido es:

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i} a_i = C_{40} = D_4 = 900 + \frac{164-120}{120} 300 = 1110\,€$$

c) Según la definición de cuantilas que se ha dado, el porcentaje pedido es el 65%, que es la diferencia entre el 90% que hay por debajo de la decila 9 y el 25% que deja a su izquierda la centila 25.

2.7 Dispersión.

El proceso de reducción estadística nos ha llevado a sintetizar todos los datos de una tabla en un solo número, al que hemos llamado promedio y con el que se pretende representar a la información que hay detrás de él. Pero para que esa medida de síntesis tenga esa validez ha de ser representativa de todos los datos contenidos en ella.

La mayor o menor representatividad de esas medidas de síntesis o promedios dependerá, fundamentalmente, del grado de concentración de todos los valores de la distribución en torno a ese promedio, cuya representatividad estamos estudiando. En el caso extremo de que todos los valores fueran iguales, la media, por ejemplo, coincidiría con uno de ellos y ésta representaría perfectamente a todos. Pero esta situación extrema nunca se da, (si se diera no tendríamos variable). Lo normal es que una variable tome más de un valor. Es en estos casos donde hay que estudiar si los promedios utilizados son representativos del conjunto de valores a los que representan.

En general, el promedio más utilizado es la media aritmética. Por esta razón nos vamos a detener en definir medidas que cuantifiquen el grado de representatividad de la media. Así, diremos que la media es representativa de una distribución si los valores de la misma están muy próximos a ella. Por el contrario, si esos valores estuvieran muy dispersos o alejados, diríamos que la media no es muy representativa.

Cuando se habla de la representatividad de un promedio es más frecuente utilizar el término dispersión que el de concentración, razón por al cual en adelante usaremos esta expresión.

Para medir la dispersión de un promedio nos basaremos en el concepto de distancia o desviaciones existentes entre los valores de la distribución y el promedio que estamos utilizando. Cuanto mayores sean estas distancias o desviaciones mayor será la dispersión y menor será esta si las distancias son pequeñas.

Las medidas de dispersión se pueden clasificar en dos categorías según que sean medidas absolutas o relativas. Dentro de las primeras las más simples son las que se basan en el recorrido de la variable (recorrido y recorridos intercuartílicos). Las más elaboradas son las que se definen en términos de distancias o desviaciones de los valores de la variable respecto de algún promedio concreto (desviación estándar, variancia, desviación media). A su vez la medidas de dispersión relativas más utilizadas son el coeficiente de apertura o disparidad y el coeficiente de variación.

2.8 Medidas de dispersión absoluta.

2.8.1 Variancia y desviación estándar.

La variancia (S^2) es la medida de dispersión más conocida y utilizada de todas cuantas puedan definirse. La misma se basa en la idea de promediar las desviaciones respecto de la media aritmética. Pero el promedio que se utiliza para obtener esa medida de dispersión no es la media aritmética de las desviaciones, pues ésta sería siempre nula, por la primera propiedad de la media aritmética, que nos habla de que la suma de las desviaciones respecto de la media es siempre cero. Este inconveniente se resuelve calculando no la media aritmética de las desviaciones sino la media cuadrática, que no es otra cosa que la media de las desviaciones respecto de la media al cuadrado. A este promedio de las desviaciones se le conoce como variancia.

La variancia[2] se define como:

$$S^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 n_i}{N} \qquad (2.11)$$

Cuanto mayor sea la variancia mayor será la dispersión de la variable y menos representativa será la media como promedio de todos los valores y viceversa. Los valores que puede tomar la variancia, dada la forma en que está definida (media cuadrática), serán siempre mayores o iguales que cero. El caso extremo se dará cuando todos los valores de la variable sean iguales, pues en tal caso todas las desviaciones serán nulas. Salvo esta situación extrema, y carente de interés para la Estadística, las desviaciones serán siempre mayores que cero y, en consecuencia, la variancia también lo será. Finalmente hay que señalar que la variancia viene expresada en las unidades de medida de la variable al cuadrado, lo que limita su interpretación.

Para salvar este último inconveniente se define la desviación estándar (S) o desviación típica, que no es más que la raíz cuadrada de la variancia tomada con signo positivo. Esta forma de definirla hace que su unidad de medida sea la misma que la de la variable.

2.8.1.1 Propiedades de la variancia.

1ª La variancia siempre será mayor o igual que cero.

Esta propiedad es evidente dada la definición de la variancia como media de sumas de desviaciones al cuadrado.

2ª La media cuadrática de las desviaciones de una variable respecto de una constante cualquiera se hace mínima cuando esa constante es la media aritmética, es decir, cuando se trabaja con la variancia. (Teorema de König).

[2] La variancia se ha definido como una media cuadrática y, en consecuencia, el denominador de la misma es el tamaño de la población. Sin embargo, resulta habitual que en los paquetes estadísticos, tales como el SPSS y otros, calculen la variancia dividiendo por (N-1). A ese cociente, para diferenciarlo del que se obtiene cuando el denominador es N, se le llama habitualmente como cuasivariancia. La explicación por la cual se utiliza como denominador N-1 en lugar de N hay que buscarla en el campo de la Inferencia Estadística, mientras que el contenido de este libro está dedicado a la Estadística Descriptiva.

La demostración de esta propiedad es inmediata a partir de la propiedad 3ª de la media aritmética.

3ª La variancia no cambia si a los valores de la variable se les suma o resta una constante (cambios de origen).

Para demostrar esta propiedad arrancaremos de la variancia de una variable *X* a la que llamaremos S_x^2 y a partir de ella definimos otra variable *Y* en la forma *Y = X + C*, donde *C* es una constante. Con esta notación vamos a obtener la variancia de *Y*.

$$S_Y^2 = \frac{\sum_{i=1}^{k}\left(y_i - \bar{y}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{k}\left(\left(x_i + C\right) - \left(\bar{x} + C\right)\right)^2 n_i}{N} = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 n_i}{N} = S_X^2$$

4ª Si los valores de la variable los multiplicamos por una constante, la variancia queda multiplicada por el cuadrado de esa constante. (Cambio de escala).

Al igual que para la propiedad anterior, sea *X* una variable con variancia S_x^2 . Definamos ahora *Y = XC,* donde *C* es una constante cualquiera. Entonces:

$$S_Y^2 = \frac{\sum_{i=1}^{k}\left(y_i - \bar{y}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{k}\left(\left(x_i C\right) - \left(\bar{x} C\right)\right)^2 n_i}{N} = \frac{C^2\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 n_i}{N} = C^2 S_X^2$$

Anteriormente hemos definido la variancia mediante la expresión (2.11). Sin embargo, para calcularla se suele recurrir al siguiente desarrollo de la misma:

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2$$

el cual facilita[3] considerablemente su obtención.

Esta expresión es válida tanto para distribuciones no agrupadas como para las agrupadas, con la diferencia que para el primer caso da un valor exacto de la variancia de la distribución, mientras que para el segundo solo da un valor aproximado debido a los errores de agrupamiento que se comenten en este tipo de distribuciones, como ya se ha señalado en otras ocasiones. Para este segundo tipo de distribuciones x_i es la marca de clase o valor central del intervalo y se supone, como se ha hecho en otras circunstancias, que la frecuencia del intervalo se distribuye de manera uniforme entre todos los valores del mismo, pues el agrupamiento de valores en intervalos nos impide saber cual es la frecuencia verdadera de cada uno de esos valores, con la consiguiente pérdida de información.

Ejemplo 13. *Obtener la variancia de la distribución de salarios del Ejemplo 11.*

Salarios (x_i)	***Asalariados (n_i)***	***Marca de clase (x_i)***	***$x_i n_i$***	***$x_i^2 n_i$***
Menos de 500	*50*	*250*	*12500*	*3125000*
De 500 a 900	*70*	*700*	*49000*	*34300000*
De 900 a 1200	*120*	*1050*	*126000*	*132300000*
De 1200 a 1800	*100*	*1500*	*150000*	*225000000*
De 1800 a 2700	*50*	*2250*	*112500*	*253125000*
De 2700 a 5000	*20*	*3850*	*77000*	*296450000*
Total	***410***		***527000***	***944300000***

En primer lugar calcularemos la media como paso previo para obtener la variancia.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{527000}{410} = 1285{,}4 \quad €$$

[3] En este caso, como en otros, se está hablando de "facilitar" el cálculo de ciertas medidas estadísticas, cuando, en realidad, hoy en día la accesibilidad y disponibilidad de paquetes estadísticos o de hojas de cálculo hace que esas preocupaciones resulten casi una cuestión del pasado.

A continuación obtendremos la variancia de esta variable:

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{944300000}{410} - (11285{,}4)^2 = 651005{,}4 \ €^2$$

Para distribuciones campaniformes y no fuertemente asimétricas se cumple que el intervalo definido por :

$$\overline{X} \pm 2S$$

contiene aproximadamente el 95% de las observaciones y el definido como:

$$\overline{X} \pm 3S$$

contiene aproximadamente el 99% de las observaciones.

Estos intervalos pueden utilizarse a su vez como medias de dispersión.

Para el ejemplo anterior, el primer intervalo contiene el 95,5% mientras que el segundo contiene el 97,25%.

2.8.2 Desviación media.

Cuando se introdujo el concepto de variancia se señaló que la media aritmética de la desviaciones no servía como indicador de dispersión porque siempre era cero. Ante tal situación se utilizó la media cuadrática para evitar que las desviaciones, positivas y negativas, se compensaran y su suma fuera nula. Otra forma de evitar esa compensación es sumar el valor absoluto de las desviaciones y dividirlo por el número de sumandos. A esta media aritmética se le conoce como desviación media. La desviación se puede definir respecto de la media, la mediana o la moda. Según el promedio que se utilice se tendría:

$$D_x = \frac{\sum_{i=1}^{N} \left| x_i - \bar{x} \right| n_i}{N} \qquad (2.12) \qquad D_{Me} = \frac{\sum_{i=1}^{N} \left| x_i - Me \right| n_i}{N} \qquad (2.13)$$

$$D_{Mo} = \frac{\sum_{i=1}^{N} \left| x_i - Mo \right| n_i}{N} \qquad (2.14)$$

2.8.3 Otras medidas absolutas de dispersión.

Además de las mediadas de dispersión señaladas con anterioridad, también se puede utilizar el recorrido, definido como $R_e = x_n - x_1$, y el recorrido intercuartílico que se define como $R_I = Q_3 - Q_1$. Estas medidas, como cualquier otro tipo de recorrido, tienen menor utilidad para medir la dispersión que las vistas con anterioridad. Además, si lo que se busca son medidas de dispersión que indiquen la representatividad de los promedios, éstas no nos sirven pues en su definición no están implicados.

Ejemplo 14.- *La notas de un examen realizado por 200 alumnos son las se dan en la tabla siguiente:*

Puntuación (x_i)	Alumnos (n_i)
2	4
3	6
4	10
4.5	14
5	28
6	35
6.5	30
7	20
7.5	20
8	15
8.5	10
9	3
9.5	3
10	2
	200

Obtenga la medidas de dispersión absoluta más habituales para esta distribución.

Para la obtención de estas medidas lo primero que debe hacerse es ampliar el contenido de la tabla anterior en la forma que se indica a continuación:

Puntuación (x_i)	**Alumnos (n_i)**	$x_i n_i$	N_i	$x_i^2 n_i$	$\lvert x_i - \bar{x}\rvert n_i$	$\lvert x_i - Me\rvert n_i$	$\lvert x_i - Mo\rvert n_i$
2	4	8	4	16	16,9	18	16
3	6	18	10	54	19,3	21	18
4	10	40	20	160	22,2	25	20
4.5	14	63	34	283,5	24,1	28	21
5	28	140	62	700	34,2	42	28
6	35	210	97	1260	7,8	17,5	0
6.5	30	195	127	1267,5	8,3	0	15
7	20	140	147	980	15,6	10	20
7.5	20	150	167	1125	25,6	20	30
8	15	120	182	960	26,7	22,5	30
8.5	10	85	192	722,5	22,8	20	25
9	3	27	195	243	8,3	7,5	9
9.5	3	28,5	198	270,75	9,8	9	10,5
10	2	20	200	200	7,6	7	8
	200	**1244,5**		**8242,3**	**249,2**	**247,5**	**250,5**

Con toda esta información se tiene que:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{1244{,}5}{200} = 6{,}2$$

Mo = 6 *Me = 6,5*

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{8242{,}3}{200} - (6{,}2)^2 = 2{,}49$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{N} \left| x_i - \bar{x} \right| n_i}{N} = \frac{249{,}2}{200} = 1{,}246$$

$$D_{Me} = \frac{\sum_{i=1}^{N} \lvert x_i - Me \rvert n_i}{N} = \frac{247{,}5}{200} = 1{,}238$$

$$D_{Mo} = \frac{\sum_{i=1}^{N} |x_i - Mo| n_i}{N} = \frac{250,5}{200} = 1,253$$

2.9 Medidas relativas de dispersión.

Todas las medidas de dispersión definidas anteriormente tienen un inconveniente en común. Todas ellas vienen expresadas en la misma unidad de medida que la variable (o al cuadrado, caso de la variancia) y todas ellas son sensibles a los cambios de escala. Esto hace que no sea posible realizar comparaciones entre la dispersión de dos distribuciones distintas. Pero, incluso aunque las unidades fueran las mismas, bastaría con que los promedios sean muy distintos para que esa comparación tampoco fuera posible.

Una forma de eliminar esos problemas es haciendo que las mismas sean adimensionales. El resultado es un conjunto de medidas de dispersión relativas. Dentro de esta categoría se engloban el coeficiente de disparidad o apertura y el coeficiente de variación de Pearson.

El primero se define como $A = x_n / x_1$. Tiene a su favor que es muy fácil de calcular. Por el contrario, presenta el grave inconveniente de que en su definición no recoge ningún promedio y que, además, está fuertemente influido por los valores extremos de la distribución, sin tener en cuenta cual es la dispersión de los demás valores.

El segundo se define como el cociente de la desviación estándar entre la media aritmética.

$$CV = \frac{S}{\bar{x}} \qquad (2.15)$$

Este coeficiente es adimensional y será siempre mayor o igual que cero[4]. La gran limitación del mismo viene por el hecho de que cuando la media es cero entonces carece de sentido. También tiene el inconveniente de que no es invariante ante cambios de origen, pues aunque la desviación estándar lo sea, la media no lo es. En cambio si que es invariante a cambios de escala. Otra limitación de este coeficiente se deriva de que no está acotado por la derecha, lo que no permite afirmar cuando es la media representativa o deja de serlo. Como regla aproximada se puede seguir el criterio de que si el coeficiente variación es menor de 0,5, entonces se puede hablar de que la media es representativa, mientras que es si es mayor que ese valor habría que cuestionar la representatividad de ese promedio. En cualquier caso, si *CV >1*, la media es poco o nada representativa.

Ejemplo 15. *Para dos distribuciones de renta distintas, una expresada una en miles de pesetas (X) y otra en euros (Y), se tienen los siguientes resultados:*

$$\bar{x} = 2200 \quad miles \quad de \quad ptas \quad S_x = 1000 \quad miles \quad de \quad ptas.$$
$$\bar{y} = 15000 \quad € \qquad\qquad S_y = 10000 \quad €$$

Analizar la dispersión y representatividad de las medias.

Este es una caso claro donde ni las medias ni las desviaciones estándares son comparables, pues estamos trabajando con unidades de medida distintas. Todavía si las medias hubieran sido iguales y las unidades de medida también, solo en ese caso las desviaciones estándares serían comparables y se podría determinar qué media es más representativa. Pero como se ha visto esta no es la situación. Ante estas circunstancias se hace necesario recurrir al coeficiente de variación de Pearson, que elimina los inconvenientes (unidades de media y distintas medias) señalados más arriba. Para este ejemplo los coeficientes son:

[4] Cuando se afirma que el CV es mayor o igual que cero se está dando por sentado que la media de la distribución es siempre mayor que cero (cuando la media es cero este coeficiente carece de sentido e interés). Este es el caso habitual cuando se trabajan con variables de tipo económico, aunque no siempre ha de ser así. Piénsese en variables del tipo beneficios empresariales, rendimientos de acciones y otras similares. En estos casos lo que suele hacerse es tomar el valor absoluto de la media o cuestionarse si la media aritmética es el promedio más adecuado en esa situación.

$$CV_x = \frac{S_x}{\overline{x}} = \frac{1000}{2200} = 0{,}4545 \qquad CV_y = \frac{S_y}{\overline{y}} = \frac{10000}{15000} = 0{,}6666$$

En este caso la media de la distribución expresada en pesetas es más representativa, pues la dispersión relativa de esta variable es menor.

Ejemplo 16.- *La distribución de los hogares de un determinado barrio, según el tamaño de los mismos, es la que muestra la tabla siguiente:*

Tamaño de hogares x_i	***Nº de hogares*** n_i
1	*40*
2	*70*
3	*110*
4	*90*
5	*48*
6	*42*
7	*40*
8	*35*
9	*20*
10	*5*
Total	***500***

Determínese:

1º ¿Cuál es el número medio de personas por hogar?

2º ¿Cuál es el tipo de hogar más frecuente?

3º Si solo hubiera plazas de aparcamiento para el 50% de las hogares y éstas se asignaran a las de mayor tamaño, ¿a partir de qué tamaño de hogar se le asignarían plaza de garaje?

4º Si en otro barrio el coeficiente de variación e 1, ¿en cuál de los barrios la media resulta más representativa?

1º En este apartado lo que se pide es que se calcule la media aritmética. El valor de la misma es:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{2152}{500} = 4{,}3 \text{ personas}$$

2º Ahora lo que se pide es la moda, el valor de la variable que más se repite. Como se trata de una variable discreta con los valores sin agrupar, la obtención de la moda es inmediato. El valor más frecuente para este ejemplo es Mo = 3 personas.

3º El tamaño de hogar que se nos pide es el mediano, pues sabemos que la mediana es el valor de la variable que divide a la distribución en dos partes iguales, es decir, a su izquierda queda el mismo número de observaciones que a la derecha. Este promedio es el que se corresponde con la primera frecuencia acumulada mayor o igual que 250 (N/2). Ese valor es Me = 4 personas.

4º Para responder a esta cuestión se hace necesario obtener previamente la desviación estándar de esta distribución. Como sabemos $S = \sqrt{S^2}$ *y*

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{11782}{500} - (4{,}304)^2 = 5{,}04 \text{ personas}^2$$

Por lo que S = 2,24 y el coeficiente de variación de esta distribución es

$$CV = \frac{S}{\bar{x}} = \frac{2{,}24}{4{,}3} = 0{,}52$$

que, comparado con el del otro barrio, nos lleva a que el tamaño de la familia en el barrio objeto de estudio está menos disperso, lo que hace que su media sea más representativa.

Ejemplo 17. *Responda a las cuestiones del Ejemplo 16 y compare los resultados si los datos se hubieren agregado en la forma siguiente*:

Tamaño de hogar x_i	***Nº de hogares*** n_i
De 0 a 2	*110*
De 2 a 4	*200*
De 4 a 6	*90*
De 6 a 8	*75*
De 8 a 10	*25*
Total	***500***

Ahora, al tener los valores de la variable agrupados en intervalos, se hace necesario calcular las marcas de clase con vistas a obtener tanto la media con la desviación estándar. Por ello lo primero que es conveniente hacer es ampliar la tabla con otras columnas adicionales.

Tamaño de hogar	n_i	x_i	$x_i\, n_i$	N_i	$x_i^2 n_i$
De 0 a 2	*110*	*1*	*110*	*110*	*110*
De 2 a 4	*200*	*3*	*600*	*310*	*1800*
De 4 a 6	*90*	*5*	*450*	*400*	*2250*
De 6 a 8	*75*	*7*	*525*	*475*	*3675*
De 8 a 10	*25*	*9*	*225*	*500*	*2025*
Total	***500***		***1910***		***9860***

A partir de esta tabla se tiene que:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{1910}{500} = 3{,}82 \text{ personas}$$

Como puede observarse, el nuevo valor de la media no coincide con el anterior. La diferencia es lo que se conoce como error de agrupamiento. El valor verdadero es 4,3. Sin embargo, cuando se agrupa se pierde información y el coste de esa pérdida es cometer un error. En este ejemplo es fácil ver donde está la naturaleza de este error. Basta con fijarse en la columna tercera de ésta última tabla para comprobar que los hogares de tamaño 1 no son 110, sino 40, o que los de tamaño 3 no son 200 y, así, sucesivamente.

Estos resultados muestran lo delicado que resulta agrupar los valores de las variables discretas y más cuando el número de los mismos no es muy elevado, como es este caso.

Esta agrupación nos lleva a que los demás promedios y las medidas de dispersión que se van a calcular se vean afectadas también por el error de agrupamiento, como será fácil de comprobar sin más que observar los resultados que se dan a continuación.

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i = 2 + \frac{90}{110+90} 2 = 2{,}9 \text{ personas}$$

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 2 + \frac{250-110}{200} 2 = 3{,}4 \text{ personas}$$

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{9860}{500} - (3{,}82)^2 = 5{,}13 \text{ personas}^2$$

$$CV = \frac{S}{\bar{x}} = \frac{2{,}26}{3{,}82} = 0{,}59$$

2.10 Tipificación de una variable

Relacionado con los conceptos que se han definido en apartados anteriores aparece el de tipificación (estandarización o normalización) de los valores de una variable. La tipificación no es una medida de dispersión ni un promedio. Se trata de un procedimiento que facilita la comparación entre los valores de dos distribuciones distintas.

Se dice que *una variable está tipificada, normalizada o estandarizada cuando a sus valores se les resta su media aritmética y se les divide por su desviación estándar*. El

resultado de esa operación es otra variable que tiene de media cero y de variancia la unidad. A las variables tipificadas se les representa habitualmente por la letra Z, de forma que si X es una variable con media $\bar{x}$ y con desviación estándar S_x, entonces, con la definición dada más arriba, se tiene que:

$$Z = \frac{X - \bar{x}}{S_x} \qquad (2.16)$$

La utilidad de la tipificación es doble. Por un lado nos lleva a una distribución muy especial, aquella en la que su media vale cero y su variancia es la unidad.

En segundo lugar, y como se ha indicado más arriba, permite realizar comparaciones entre valores de distintas distribuciones cuando éstas tienen medias y variancias diferentes.

Para dejar más clara esta utilidad se podría recurrir al siguiente ejemplo. Imaginemos que deseamos comparar las notas de un alumno en dos exámenes distintos. En uno de ellos ha obtenido un 6 y la nota media de ese examen para todos los alumnos ha sido de 7, con una desviación estándar de 2. En el otro examen ha obtenido una nota de 5, siendo la media para el conjunto de todos los que se examinaron de 4,5 con una desviación estándar de 1. Ante esta situación, si nos atenemos a las dos notas, la conclusión inmediata es que ha obtenido un mejor resultado en el primer examen. Pero no siendo falsa esta conclusión, lo cierto es que no son comparables esos dos valores, pues proceden de dos poblaciones totalmente distintas, con distintas medias y variancias, por lo que no son comparables. Para que esas comparaciones puedan realizarse de forma correcta hay que homogeneizar las dos distribuciones, o lo que es igual, hay que eliminarles sus características propias y reducirlas a un único patrón. Esto se consigue tipificando las variables. En nuestro caso concreto tendríamos los siguientes resultados:

$$Z = \frac{X - \bar{x}}{S_x} = \frac{6-7}{2} = -0{,}5 \qquad Z = \frac{X - \bar{x}}{S_x} = \frac{5-4{,}5}{1} = 0{,}5$$

La conclusión a la que se llega ahora es que la nota relativa del primer examen es inferior a la del segundo, pues si bien en el primer caso obtuvo un 6, resulta que esa nota es inferior a la media del curso, mientras que en el segundo de los exámenes su nota, aunque más baja que la primera, está por encima de la media y como además la dispersión es aún más pequeña ello lleva a que muy pocos alumnos obtuvieron notas por encima de la media, lo que confirma que el resultado relativo del segundo examen es mejor que el correspondiente al primero.

Otra situación donde queda muy clara la utilidad de la tipificación es cuando se quiere comparar el poder adquisitivo de la renta de dos personas que viven en lugares distintos. Supongamos que la renta anual de un ciudadano que reside en el país A es de 20000 € anuales, siendo la renta media de ese país de 15000 € y la desviación estándar de 10000 €. En otro país B, la renta de un residente es 25000 €, mientras que la renta media de este segundo país es 30000 € y la desviación estándar es 25000 €. Con estos datos la cuestión que se plantea es saber cual de esos dos ciudadanos tiene una renta relativa más elevada. De nuevo hay que recurrir a la tipificación para que las distribuciones de renta sean comparables u homogéneas. Ahora:

$$Z_A = \frac{X_A - \overline{x_A}}{S_A} = \frac{20000 - 15000}{10000} = 0{,}5 \qquad Z_B = \frac{X_B - \overline{x_B}}{S_B} = \frac{25000 - 30000}{25000} = -0{,}2$$

de lo que se deduce que al renta relativa del residente en A es mayor que la del residente en B.

2.11 Medidas de forma. Asimetría y curtosis.

En los epígrafes anteriores hemos estudiado distintas características de una distribución. El cuarto lo hemos dedicado a estudiar la forma de una distribución y en el mismo se describieron algunos modelos de distribuciones que se repiten con cierta frecuencia. Ello nos llevó a hablar de distribuciones campaniformes, en forma de *U* y en forma de *J*. Pues bien, este epígrafe intenta ahondar en esta cuestión, pero ahora en lugar de utilizar la representación gráfica para hablar de la forma de una distribución haremos uso de ciertas medidas sintéticas. Estas medidas servirán para determinar el grado de simetría y el de apuntamiento de la distribución (curtosis). Esta segunda característica tiene un carácter

relativo en la medida que se da en relación con otra distribución que se usa como modelo de referencia (la distribución normal).

2.11.1 Simetría de una distribución.

Diremos que una distribución es simétrica respecto de un determinado eje si existe el mismo número de valores a cada lado de ese eje, equidistantes respecto del mismo dos a dos y tales que para cada par de valores equidistantes a dicho eje tengan la misma frecuencia. En caso contrario se dice que la distribución no es simétrica. El eje de simetría que suele utilizarse como referente principal es el valor de la media aritmética.

Si tomamos la media como eje de simetría y la distribución es simétrica, entonces las desviaciones de los valores de la distribución con respecto a la media serán positivas y negativas y habrá tantas positivas como negativas, de tal manera que su suma será nula. Pero esa suma será siempre nula, incluso en aquellos casos de distribuciones no simétricas (recordemos que la suma de las desviaciones con respecto a la media es cero, incluso cuando la distribución no es simétrica). En consecuencia esta suma de desviaciones no nos permite determinar si predominan las diferencias positivas o negativas, pues si las diferencias positivas fueran mayores que las negativas ello significaría que la mayor parte de la distribución se encuentra a la derecha de la media (distribución asimétrica a la derecha o positiva), mientras que si fueran las diferencias negativas las que predominaran sobre las positivas entonces nos encontraríamos con que la mayor parte de los valores de la distribución estaría en la cola izquierda de la misma, a la izquierda de la media (distribución asimétrica a la izquierda o negativa).

Vemos pues que, una medida que se base en la suma de esas diferencias y que mantenga el signo de las que dominan será un buen indicador de la presencia o no de simetría.

Si las diferencias respecto de la media las elevamos a una potencia par evitamos que su suma sea nula, pero el signo de esas diferencias será siempre positivo, con lo cual este tipo de medias no nos sirve. En cambio si la potencia que utilizamos es impar entonces lo que conseguimos es que la suma no se anule y que además las desviaciones mantengan su signo. De todas las posibles potencias impares la más simple es la tercera. Así nuestro coeficiente de asimetría vendría dado por:

$$m_3 = (S_i(x_i - x)^3 n_i)/N \qquad (2.17)$$

de forma que si:

$m_3 = 0$ la distribución es simétrica
$m_3 > 0$ la distribución es asimétrica a la derecha o positiva
$m_3 < 0$ la distribución es asimétrica a la izquierda o negativa

Esta medida de asimetría viene expresada en las unidades de la variable al cubo y además no es invariante a los cambios de escala. Estos son dos inconvenientes que se pueden salvar fácilmente si la dividimos por el cubo de la desviación estándar. A la medida resultante se le conoce como coeficiente de asimetría de Fisher. El mismo viene dado por:

$$\gamma_1 = m_3/S^3 \qquad (2.18)$$

Como *S* es siempre positiva, resulta que el signo de γ_1 es el m_3 con lo que:

$g_1 = 0$ la distribución es simétrica
$g_1 > 0$ la distribución es asimétrica a la derecha o positiva
$g_1 < 0$ la distribución es asimétrica a la izquierda o negativa

La principal limitación de este coeficiente es que para distribuciones simétricas el mismo vale siempre cero, pero el recíproco no es siempre cierto. En estas circunstancias habría que recurrir a la representación gráfica.

Para el caso de distribuciones campaniformes unimodales y simétricas sabemos que la media y la moda coinciden. Pero sin en distribuciones campaniformes unimodales esos dos promedios no coincidieran, entonces ello sería indicativo de que la distribución no es simétrica. Basándose en este principio K. Pearson definió el siguiente coeficiente de asimetría:

$$A = \frac{\bar{x} - Mo}{S} \qquad (2.19)$$

cuya interpretación es la siguiente:

A = 0 la distribución es simétrica
A > 0 la distribución es asimétrica a la derecha o positiva
A < 0 la distribución es asimétrica a la izquierda o negativa

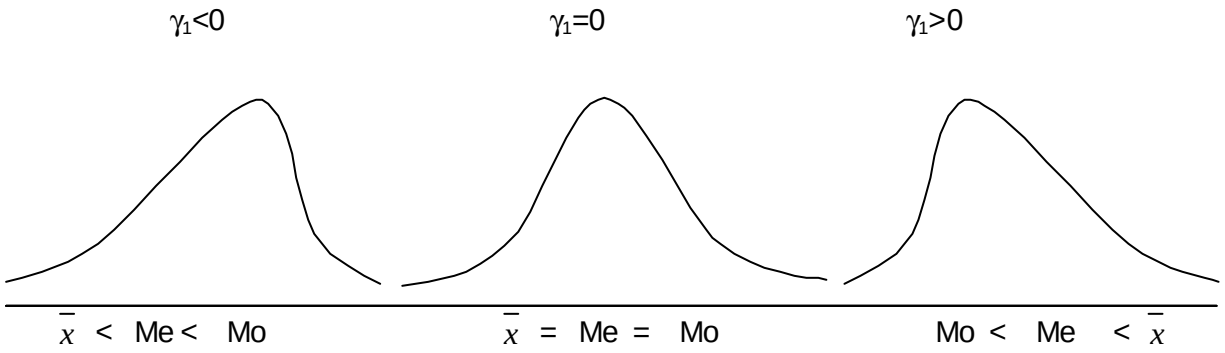
Para este tipo de distribuciones campaniformes, si además no son muy asimétricas, se cumple la siguiente relación aproximada:

$$\bar{x} - Mo \cong 3\left(\bar{x} - Me\right) \qquad (2.20)$$

por lo que también, de forma aproximada, se puede decir que:

$$A \cong \frac{3\left(\bar{x} - Me\right)}{S} \qquad (2.21)$$

Figura 12. Análisis gráfico de la simetría.



2.11.2 Apuntamiento de una distribución (Curtosis)

Esta característica de una distribución es la que tiene un menor grado de interés de todas las que se han considerado hasta el momento. La misma hace referencia al mayor o menor grado de apuntamiento de una distribución y se define en términos de la distribución normal. Para la distribución normal se verifica que $m_4/S^4 = 3$, donde:

$$m_4 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^4 n_i}{N} \qquad (2.22)$$

A partir de esta relación se define el coeficiente de curtosis o apuntamiento de la forma siguiente:

$$g_2 = (m_4/S^4) - 3 \qquad (2.23)$$

de forma que si:

- $g_2 = 0$ distribución mesocúrtica (normal)
- $g_2 > 0$ distribución leptocúrtica
- $g_2 < 0$ Distribución platicúrtica

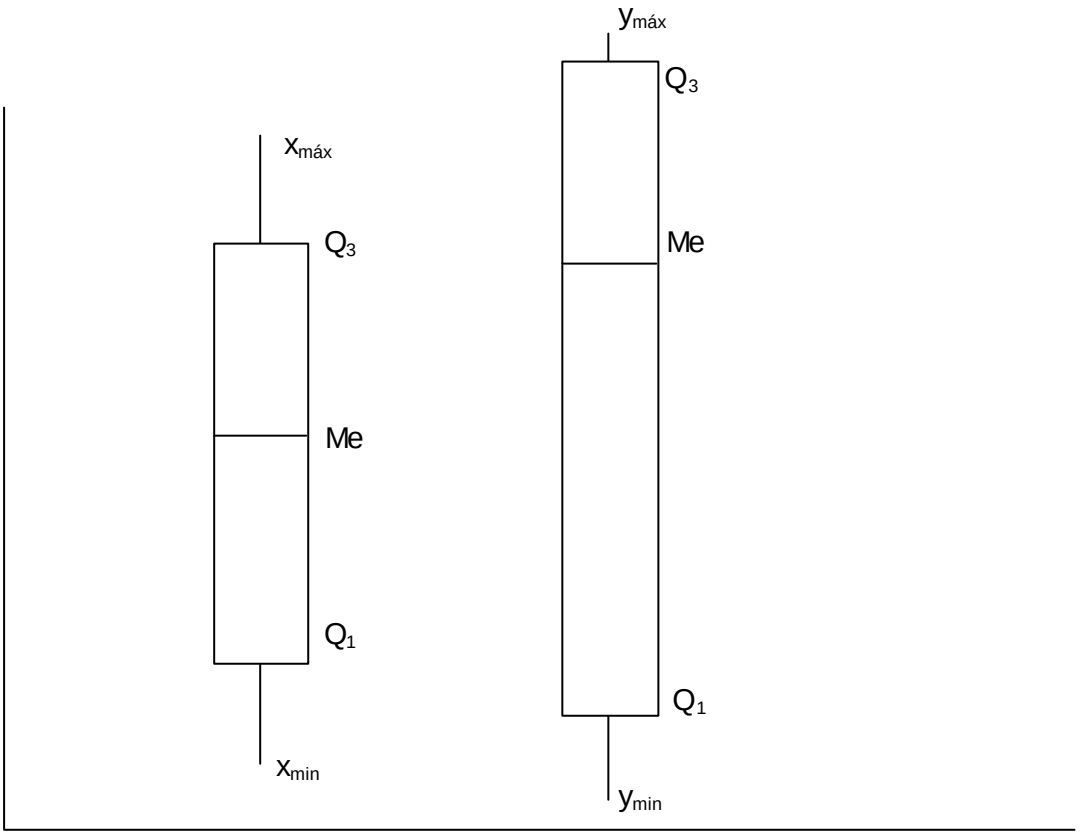
2.12 Análisis gráfico de la dispersión y la asimetría.

Estas dos características de una distribución las hemos estudiado hasta ahora de forma separada mediante medidas específicas o inspeccionando la representación gráfica de la distribución. Si embargo es posible, mediante lo que se conoce como diagramas de caja, realizar un análisis simultáneo de ambas características. Estos diagramas se basan en el recorrido de la distribución, la mediana y los valores de la primera y tercera cuartila.

De forma genérica, en la Figura 13 se han representado dos de estos diagramas, uno para la distribución de una variable *X* y otro para la de una variable *Y* cualesquiera. Para cada una de esas distribuciones se han representado los valores extremos, las medianas y la primera y tercera de las cuartilas. El resultado son dos representaciones gráficas que nos informan de lo siguiente: al variable *X* tiene menor dispersión que la *Y*, pues su

recorrido es menor y además el cincuenta por ciento de las observaciones centrales (las comprendidas entre Q_1 y Q_3) se concentran en un intervalo menor. Por otro lado, la distribución de X es casi simétrica, pues la mediana está en centro de los recorridos intercuartílico y total, mientras que en el caso de Y se encuentra más próxima a los valores grandes, lo que nos indica que se trata de una distribución asimétrica a la izquierda.

Figura 13. Diagramas de caja.



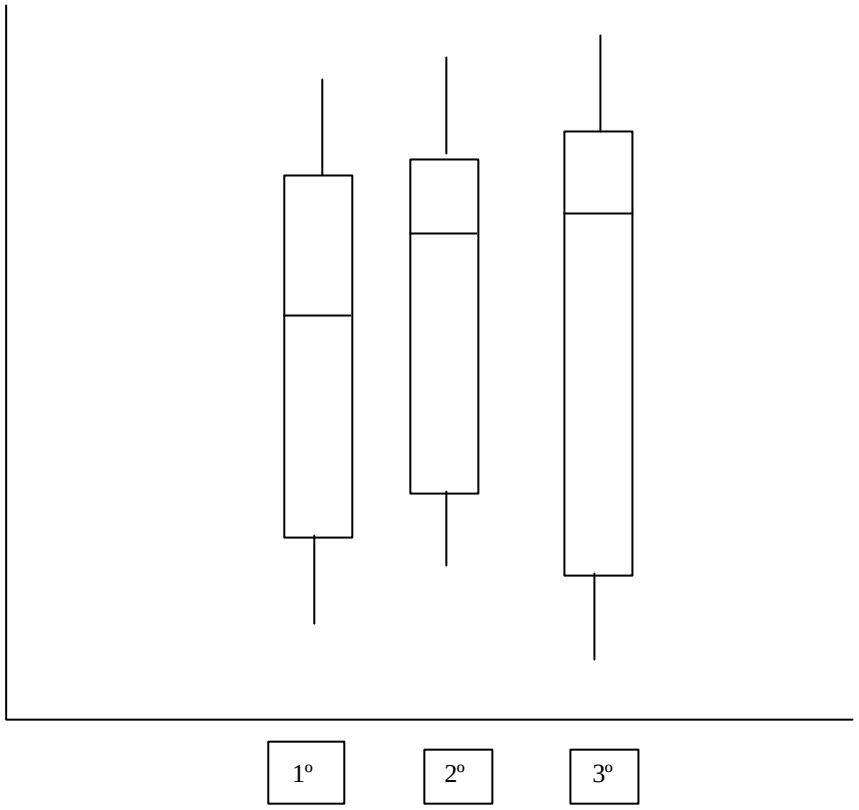
***Ejemplo 18**. En la tabla siguiente se dan los resultados de un examen en tres cursos distintos. Obtener los correspondientes diagramas de caja y estudiar la dispersión y simetría de esas distribuciones.*

Primero	*Segundo*	*Tercero*
4,7	5,6	4,3
5,2	5,9	4,8
5,2	5,9	5,0
5,7	6,1	5,5
6,3	6,7	6,1
6,4	6,9	6,7
6,9	7,3	7,2
7,1	7,6	7,8
7,2	7,6	8,0
7,2	8,0	8,0
7,8	8,3	8,3
8,1	8,3	8,5
8,1	8,4	8,9
8,6	9,0	9,1
9,1	9,4	9,7

A partir de estos datos se obtuvieron los siguientes resultados, los cuales nos permiten construir los correspondientes diagramas de caja.

	Primero	*Segundo*	*Tercero*
Mín.	4,7	5,6	4,3
Máx.	9,1	9,4	9,7
Q_1	5,7	6,1	5,5
Me	7,1	7,6	7,8
Q_3	8,1	8,3	8,5

Estos gráficos muestran como la tercera distribución es asimétrica a la izquierda y tiene una dispersión mayor que las otras dos. Por el contrario la primera es casi simétrica mientras que la segunda es la menos dispersa, pues el cincuenta por ciento de las observaciones centrales se concentra en un recorrido pequeño.



2.13 Momentos de una distribución.

Hasta ahora hemos venido hablando de distintas características de una distribución. Las mismas se han cuantificado mediante una serie de medidas, tales como la media aritmética, la variancia y otras más. Pues bien, algunas de ellas son casos particulares de lo que, genéricamente, se conoce como momentos. Estos son medidas que se obtienen mediante la información numérica que suministra una variable y sirven para caracterizar, de forma única, a una distribución. Tanto es así que si dos distribuciones tienen idénticos momentos, entonces se trata de la misma distribución.

Los momentos se agrupan en dos categorías. Los momentos con respecto al origen y los momentos con respecto a la media o centrales. Conceptualmente son idénticos,

diferenciándose solo en el origen que se tome como referencia. De hecho, como se verá más adelante, los unos están relacionados con los otros.

Los momentos con respecto al origen se definen de la forma siguiente:

$$m_k' = \frac{\sum_{i=1}^{n} x_i^k n_i}{N} \qquad k = 0,1,2,3,4,..... \qquad (2.24)$$

Los momentos con respecto al origen que más suelen utilizarse son los de menor orden, especialmente: m_1' y m_2'. El momento de orden cero vale siempre uno, no importa cual sea la variable. El de orden uno es la media aritmética y el de orden dos es el que interviene en el cálculo de la variancia.

Los momentos con respecto a la media o centrales se definen como:

$$m_k = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^k n_i}{N} \qquad k = 0,1,2,3,4,.... \qquad (2.25)$$

Al igual que ocurre con los momentos con respecto al origen, el número de momentos con respecto a la media es infinito, pero solo unos pocos, los primeros, son los que más se usan. Los dos primeros no suelen usarse casi nunca, pero tienen un valor concreto. Así el momento con respecto a la media de orden cero (m_0) vale uno, al igual que su correspondiente con respecto al origen. El de orden uno (m_1) vale cero, pues se trata de la suma de las desviaciones con respecto a la media. El de orden dos (m_2) es la variancia de la variable. El de orden tres (m_3) se utiliza para definir el coeficiente de asimetría de Fisher. El de orden cuatro se usa para el análisis de la curtosis.

A continuación vamos a dar la relación que existe entre los momentos con respecto al origen y los momentos con respecto a la media. Para ello hay que hacer uso del desarrollo del binomio de Newton:

$$\left(x_i - \bar{x}\right)^k = \sum_{j=0}^{k} (-1)^j \binom{k}{j} x_i^{k-j} \bar{x}^j$$

Sustituyendo esta relación en la definición de momento de orden k con respecto a la media se tiene:

$$m_k = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^k n_i}{N} = \frac{1}{N} \sum_{i=1}^{n} \left[\sum_{j=0}^{k} (-1)^j \binom{k}{j} x_i^{k-j} \bar{x}^j \right] n_i = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \bar{x}^j \left(\frac{1}{N} \sum_{i=1}^{n} x_i^{k-j} n_i \right) =$$

$$= \sum_{j=0}^{k} (-1)^j \binom{k}{j} m'_{k-j} \bar{x}^j$$

Como caso particular de esta relación se tiene la siguiente:

$$S^2 = m_2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = m'_2 - \left(m'_1\right)^2$$

donde se muestra que la variancia, (momento de orden dos con respecto a la media), es igual al momento de orden dos con respecto al origen menos la media (momento de orden uno con respecto al origen) al cuadrado.

2.14 Desigualdad

En este apartado vamos a continuar hablando de concentración pero en un sentido diferente al utilizado en los epígrafes anteriores. Ahora, cuando hablemos de concentración, estaremos haciendo referencia al mayor o menor grado de igualdad o desigualdad en el reparto o distribución de una cierta magnitud. Ahora el término concentración será sinónimo de desigualdad. Las distintas medidas de concentración que pueden utilizarse, en el sentido señalado con anterioridad, son indicadores del grado de equidistribución de la variable.

Estos indicadores tienen una gran aplicación en el ámbito de la economía y especialmente dentro de lo que podría llamarse economía social, pues lo que buscan es determinar el mayor o menor grado de equidad en el reparto de variable tales como renta,

salarios, riqueza, etc, entre los preceptores de esas magnitudes. Pero estos indicadores no solo se usan para cuantificar la desigualdad en el reparto entre personas. También pueden utilizarse para estudiar el reparto en áreas geográficas, empresas, etc.

Los indicadores más habituales que se utilizan en este ámbito son la Curva de Lorenz, el Indice de Gini y el Indice de Theil. Todos ellos se basan en el siguiente principio. Sea X una variable (por ejemplo la renta) cuyos valores ordenados son los siguientes: $x_1 £ x_2 £ ..... £ x_n$, donde cada x_i es lo percibido por un sujeto. Lo que se pretende saber es si el total $S_i x_i$ (la renta total en este caso) está equitativamente distribuida. Las dos situaciones extremas que se pueden dar frente a este problema son las siguientes:

1º Que todos los perceptores tengan el mismo nivel de renta, en cuyo caso se hablaría de equidistribución o concentración mínima: $x_1 = x_2 = ... = x_n$.

2º Que *n-1* perceptores tengan renta nula y solo uno obtenga toda la renta. En este caso la concentración sería máxima.

Entre estas dos soluciones extremas pueden existir un número infinito de soluciones con distinto grado de concentración.

2.14.1 Curva de Lorenz

Para la exposición de este instrumento de concentración haremos uso de la información contenida en la Tabla 4. Los datos de esta tabla hacen referencia a una distribución genérica (x_i, n_i) cualquiera, referida a rentas, salarios, etc.

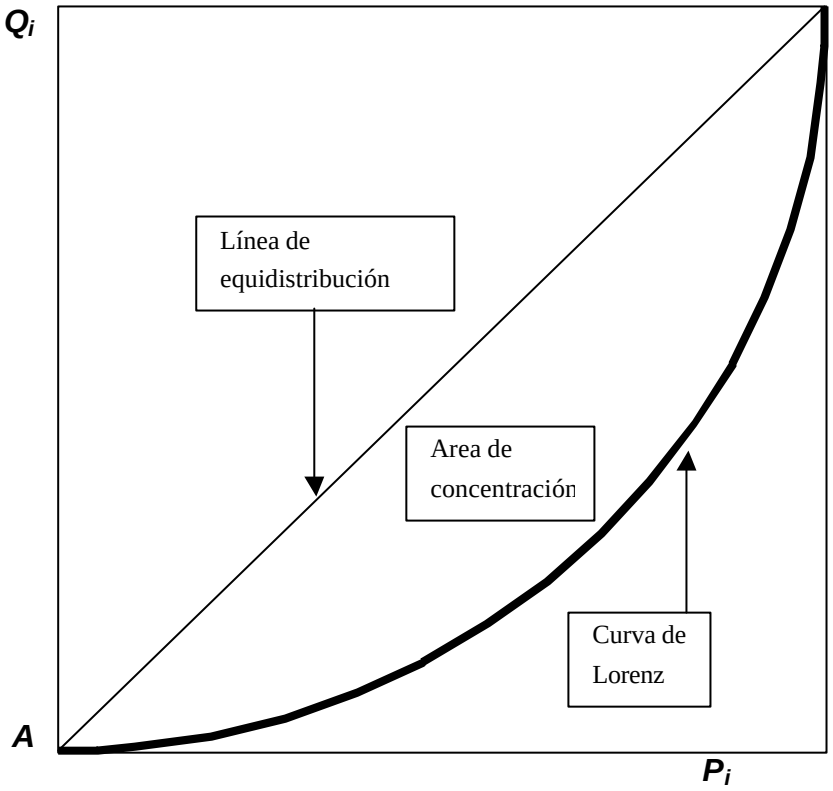
De todas las columnas de esta tabla, las que vamos a utilizar para representar la Curva de Lorenz son las encabezadas por P_i y Q_i. La representación gráfica de estos valores dará lugar a la Figura 14.

Tabla 4. Información para la Curva de Lorenz y el Índice de Gini.

x_i	n_i	$t_i = x_i\, n_i$	$p_i = (n_i/N)$	$q_i = (t_i/T)$	N_i	T_i	$P_i = (N_i/N)$	$Q_i = (T_i/T)$	$P_i Q_{i+1}$	$P_{i+1} Q_i$
x_1	n_1	$t_1=x_1\, n_1$	n_1/N	t_1/T	N_1	$T_1=t_1$	P_1	Q_1	P_1Q_2	P_2Q_1
x_2	n_2	$t_2=x_2\, n_2$	n_2/N	t_2/T	N_2	$T_2=T_1+t_2$	P_2	Q_2	P_2Q_3	P_3Q_2
x_3	n_3	$t_3=x_3\, n_3$	n_3/N	t_3/T	N_3	$T_3=T_2+t_3$	P_3	Q_3	P_3Q_4	P_4Q_3
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_i	n_i	$t_i=x_i\, n_i$	n_i/N	t_i/T	N_i	$T_i=T_{n-i}+t_i$	P_i	Q_i	P_iQ_{i+1}	$P_{i+1}Q_i$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_n	n_n	$t_n=x_n\, n_n$	n_n/N	t_n/T	$N_n=N$	$T_n=T$	$P_n=1$	$Q_n=1$	---	---
	N	$T=\Sigma\, x_i\, n_i$								

A la diagonal dibujada se le conoce como línea de equidistribución, mientras que a la línea bajo esa diagonal se le conoce como Curva de Lorenz. En realidad esa curva, cuando se obtiene como representación gráfica de un conjunto de pares de valores, como los de la Tabla 4, es una línea poligonal. Cuanto más se aleje la Curva de Lorenz de esa diagonal, mayor será el grado de concentración o desigualad en el reparto y, en consecuencia, menor la equidistribución. Esa curva siempre estará por debajo de la línea *AB,* nunca la cruzará, pues la Curva de Lorenz responde a una función monótona no decreciente. Solo coincidirá con ella cuando $P_i = Q_i$, es decir, cuando se de la situación de total igualdad en el reparto o equidistribución.

Figura 14. Curva de Lorenz



2.14.2 Índice de Gini

El índice de Gini se define como el cociente entre el área delimitada por la línea *AB* (Figura 14) y la Curva de Lorenz (área de concentración) y el área del triángulo inferior a la línea *AB*. Los valores de este índice van de cero a uno. El valor cero se alcanza cuando la Curva de Lorenz coincide con el segmento *AB*, pues en tal caso el área comprendida entre ambas es nula. Se trataría del caso de equidistribución. En este caso se habla de que no hay desigualdad en el reparto o que la concentración es nula. Todos los preceptores reciben la misma cantidad. El otro valor extremo se alcanza cuando la Curva de Lorenz está formada por los segmentos *AC* y *CB*, pues en este caso el área de concentración coincide con la del triángulo inferior. En este situación se hablaría de concentración extrema y el Indice de Gini valdría uno. Esta circunstancia se daría cuando solo un preceptor recibe toda la renta, ingreso, etc. Entre estas dos situaciones extremas y poco probables se pueden dar un conjunto infinito de valores.

La forma aproximada de calcular ese índice, que lo representaremos por *G*, se basará en la determinación del área bajo la curva. Esta área será la suma de áreas de triángulos y trapecios, pues en casos prácticos la curva de Lorenz se convierte en una línea poligonal más que en una línea continua como la dibujada en la figura anterior. Concretamente se tendría:

G = (Area de concentración)/(Area del triángulo ACB) =
=(Area de concentración)/(1/2)= 2(Area de concentración)=
= 2(1/2 – Area bajo la curva)= 1 – 2(Area bajo la curva)=

$$@ \sum_{i=1}^{n-1} P_i Q_{i+1} - \sum_{i=1}^{n-1} P_{i+1} Q_i \qquad (2.26)$$

Esta expresión de cálculo del Indice de Gini es solo una aproximación[5] a su verdadero valor, por cuanto no se trabaja en términos continuos, dado que, como ya se ha señalado antes, el área de concentración con la que se trabaja es la comprendida entre la diagonal principal (línea de equidistribución) y la poligonal que representa a la Curva de Lorenz. Pero esta no es la única forma de obtener el Indice de Gini. Otra, equivalente a la expresión anterior, es la siguiente:

$$G \cong 1 - \sum_{i=1}^{n} p_i \left[\sum_{j=1}^{i-1} 2q_j + q_i \right] = \frac{\sum_{j<i} (x_i - x_j) p_i q_j}{\bar{x}} \qquad (2.27)$$

Este índice es adimensional e invariante ante cambios de escala, pero no frente a cambios de origen, los cuales hacen que *G* se reduzca.

[5] Se dice que es una aproximación porque el área bajo la Curva de Lorenz habría que obtenerla mediante la correspondiente integral de la función que representa a la Curva. Ahora bien, si se admite como válido sustituir esa línea continua, que es la Curva de Lorenz, por la poligonal

Ejemplo 19. *Para la distribución de salarios mensuales (expresada en euros) dada en la siguiente tabla*

Salarios (x_i)	***Asalariados (n_i)***
Menos de 500	*50*
De 500 a 900	*70*
De 900 a 1200	*120*
De 1200 a 1800	*100*
De 1800 a 2700	*50*
De 2700 a 5000	*20*
Total	***410***

analizar la concentración tanto gráfica como analíticamente.

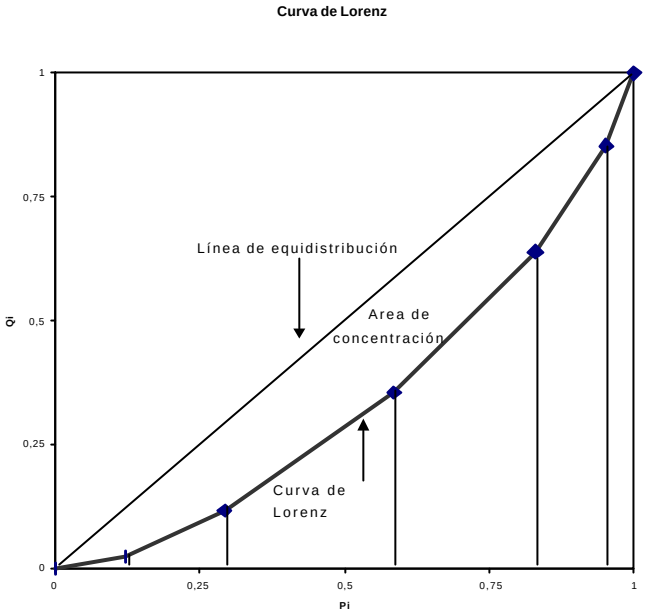
Para estudiar la concentración lo primero que se hará es completar la tabla anterior en la forma siguiente:

Marca de clase (x_i)	t_i	N_i	T_i	P_i	Q_i	$P_i\,Q_{i+1}$	$P_{i+1}Q_i$
250	*12500*	*50*	*12500*	*0,12195*	*0,02371*	*0,01423*	*0,00694*
700	*49000*	*120*	*61500*	*0,29268*	*0,11669*	*0,10413*	*0,06831*
1050	*126000*	*240*	*187500*	*0,58536*	*0,35578*	*0,37487*	*0,29504*
1500	*150000*	*340*	*337500*	*0,82926*	*0,64041*	*0,70810*	*0,60917*
2250	*112500*	*390*	*450000*	*0,95121*	*0,85388*	*0,95121*	*0,85388*
3850	*77000*	*410*	*527000*	*1*	*1*	*0*	*0*

A partir de estos datos se obtiene la Curva de Lorenz que nos permite tener una idea gráfica del grado de concentración que presenta la variable de este ejemplo. Como se aprecia, no puede decirse que concentración sea nula, pues la Curva de Lorenz se aleja de la línea de equidistribución. Además, también puede observarse por los datos de la tabla anterior que el 12% de los que menos ganan solo obtienen el 2% del total de la

correspondiente, entonces el área bajo esa poligonal obtenida mediante la expresión (2.26) es exacta, sin importar el número de valores que tome la variable.

masa salarial, mientras que el 5% de los que más ganan acaparan casi el 15% del total de los salarios. Sin embargo la concentración no es demasiado elevada.



Pero ni el gráfico anterior ni los datos elaborados de la tabla nos permiten tener una idea más precisa del grado de concentración de esta variable. Para ello hay que recurrir al cálculo del Indice de Gini. En este caso, el valor del mismo es:

$$G = \sum_{i=1}^{n-1} P_i Q_{i+1} - \sum_{i=1}^{n-1} P_{i+1} Q_i = 2{,}1526 - 1{,}8334 = 0{,}3192$$

Se trata de un valor más próximo a cero que a uno. En consecuencia, la situación no es de reparto igualitario pero tampoco puede hablarse de que la concentración sea elevada.

2.14.3 Índice de Theil.

Este indicador de desigualdad en el reparto de una magnitud entre distintas unidades preceptoras o de asignación (el reparto puede tener lugar entre personas (rentas), empresas (cuotas de mercado), unidades espaciales (provincias o regiones), etc.) fue introducido inicialmente como una medida de entropía dentro del contexto de la teoría de la información. La entropía sirve para medir el grado de desorden en un sistema (un sistema desordenado sería equivalente a otro en el que cada uno de los componentes del

mismo no están “equilibrados”) y también para comparar situaciones distintas, como veremos más adelante.

El índice de Theil se define, inicialmente, en términos de las probabilidades de los distintos valores de un distribución. Sin embargo, esas probabilidades pueden aproximarse por las frecuencias relativas observadas para esos valores o simplemente por un conjunto de proporciones, con la únicas condiciones de que sean no negativas y que su suma sea igual a la unidad. Pues bien, si nos situamos en este último contexto, podemos imaginar la situación de un conjunto de *N* empresas que se dedican a producir el mismo bien. Cada una de estas empresas tiene su propia cuota de mercado, siendo p_i la de la empresa *i-ésima*. Para este caso, el Índice de Theil viene dado por:

$$H = \sum_{i=1}^{N} p_i \log \frac{1}{p_i} \qquad (2.28)$$

Este índice está acotado y sus valores extremos son *0* y *log N.* El valor cero tiene lugar cuando la cuota de mercado de todas las empresas vale cero, salvo la de una que es la unidad (situación de monopolio). En este caso se dice que la concentración es máxima. No hay igualdad en el reparto. En general esta situación de concentración máxima se da cuando $p_i = 1$ y $p_j = 0$ para todo *j* distinto de *i*. En cambio si cada empresa tiene la misma cuota de mercado (competencia perfecta), es decir $p_i = p_j$, la concentración es mínima y se habla de reparto igualitario. En esta situación el valor del Índice de Theil es *log N.* Así pues cuanto mayor es el valor del índice menor desigualdad en el reparto y en la medida que se acerque a cero se puede hablar de mayor concentración.

Ahora bien, como puede observarse, el valor de este índice depende de *N* en su extremo superior, por lo que no sirve para realizar comparaciones cuando el tamaño de la población es distinto. En estos casos se podría recurrir a dividir el índice por log *N*. Si se procede de esta forma, los valores extremos serían siempre cero y la unidad.

El problema de la comparación puede resolverse como acaba de indicarse o bien realizando otro planteamiento. Supongamos que se tienen dos distribuciones distintas, *X* e *Y*, pero con un mismo tamaño poblacional. Si representamos por p_i a las participaciones

de los valores de *X* en su distribución y por q_i a las de *Y* en la suya, entonces el Índice de Theil para este caso viene dado por la expresión:

$$I(q:p) = \sum_{i=1}^{N} q_i \log \frac{q_i}{p_i}$$

Este indicador tomará el valor cero cuando $p_i = q_i$, es decir, cuando la igualdad o desigualad en el reparto en las dos distribuciones es idéntica. Además puede demostrarse que cuando esas proporciones no son iguales el indicador es siempre mayor que cero pudiendo hacerse infinito cuando $q_i > p_i = 0$.

Ejemplo 20. *En una determinado país existen 6 empresas que se dedican a la distribución de energía eléctrica siendo sus cuotas de mercado (proporción de energía vendida por la empresa i-ésima respecto del total de energía vendida) las que se recogen en la siguiente tabla:*

Empresa	*Cuota de mercado (p_i)*
A	*0,20*
B	*0,15*
C	*0,10*
D	*0,05*
E	*0,05*
F	*0,45*

Estudiar el grado de concentración empresarial de este sector para ese país.

En este caso, aunque también sería aplicable el índice de Gini, vamos a utilizar el de Theil, para lo cual no es necesario ordenar los valores con ningún criterio. Lo primero que debe hacerse es ampliar la tabla con aquellas columnas adicionales que recojan todos los cálculos necesarios para la obtención del índice.

Empresa	*Cuota de mercado (p_i)*	*$p_i \log(1/p_i)$*
A	*0,20*	*0,3219*
B	*0,15*	*0,2846*
C	*0,10*	*0,2303*
D	*0,05*	*0,1498*
E	*0,05*	*0,1498*
F	*0,45*	*0,3593*

Con estos datos resulta que:

$$H = \sum_{i=1}^{N} p_i \log \frac{1}{p_i} = 1{,}4956$$

que se aleja del valor cero y se aproxima bastante a log (6) = 1,7918.

Ejemplo 21. *En 1999 la distribución de la población y el VAB por provincias en Andalucía era el que se da en la tabla siguiente. Realizar un análisis de la concentración de la población y la renta en Andalucía.*

Provincias	*VAB (millones ptas)*	*Población*
Almería	*890462*	*512843*
Cádiz	*1685973*	*1119802*
Córdoba	*1132683*	*768676*
Granada	*1125698*	*813061*
Huelva	*712562*	*457507*
Jaén	*946873*	*649662*
Málaga	*2051469*	*1258084*
Sevilla	*2752314*	*1725482*

En este caso se podría plantear calcular el Índice de Theil para cada una de las distribuciones y compararlo. Pero con esto no se consigue el objetivo que se busca y que no es otro que determinar si la distribución de la población y la producción siguen el mismo patrón, es decir, si la producción per capita es parecida o si por el contrario la producción no tiene lugar donde está la población. Para conseguir este objetivo se puede utilizar el Índice de Theil dado por:

$$I(q:\ p) = \sum_{i=1}^{N} q_i \log \frac{q_i}{p_i} = 0{,}0018$$

Como puede apreciarse el valor del índice es muy bajo, lo que significa que la producción per capita es muy parecida entre todas las provincias, es decir que la concentración de la población es similar a la de la producción. Pero la información que contiene la tabla que se ha elaborado para calcular este índice nos advierte que hay cuatro provincias que contribuyen o concentran, proporcionalmente, más producción que población. En este caso se trata de Almería, Huelva, Málaga y Sevilla.

Provincias	VAB (millones ptas)	Población	p_i	q_i	$q_i \log(q_i/p_i)$
Almería	890462	512843	0,07020	0,07881	0,00912
Cádiz	1685973	1119802	0,15329	0,14922	-0,00400
Córdoba	1132683	768676	0,10522	0,10025	-0,00485
Granada	1125698	813061	0,11132	0,09963	-0,01109
Huelva	712562	457507	0,06262	0,06306	0,00044
Jaén	946873	649662	0,08893	0,08380	-0,00497
Málaga	2051469	1258084	0,17221	0,18157	0,00960
Sevilla	2752314	1725482	0,23620	0,24360	0,00752

1-. La distribución del presupuesto semanal en alimentación de un conjunto de 265 familias, expresado en euros, es el que figura en la tabla siguiente:

Presupuestos L_{i-1} - L_i	**Familias** n_i
80-100	10
100-110	35
110-115	40
115-120	45
120-130	55
130-150	30
150-170	20
170-210	15
210-270	10
270-360	5
Total	**265**

A partir de esa información:

1º Diga cual es la población, los elementos, la característica observada, el tipo de variable.
2º Represente gráficamente esta variable.
3º Obtenga la media, mediana, moda, variancia, desviación estándar, coeficiente de variación, coeficiente de asimetría, coeficiente de curtosis.
4º Determine el recorrido intercuartílico, interdecílico e intercentílico.
5º Determine el número de familias con un presupuesto inferior a la media.
6º Determine el porcentaje de familias con un presupuesto comprendido entre la media menos dos veces la desviación estándar y la media más dos veces la desviación estándar.
7º Obtenga la curva de concentración de Lorenz.
8ª Determine el Indice de Gini.
9ª Responda a todos los apartados anteriores si la unidad monetaria fuera la peseta.

2.- La distribución de los salarios semanales en una determinada industria, expresada en euros, es la que figura en la tabla siguiente:

Salarios $L_{i-1} - L_i$	**Asalariados** n_i
Menos de 90	2145
90-120	1520
120-150	840
150-180	955
180-210	1110
210-240	2342
240-300	610
300-600	328
Más de 600	150
Total	**10000**

A partir de esa información y sabiendo que la nomina total de esa industria asciende a 1765950 €

1º Diga cual es la población, los elementos, la característica observada, el tipo de variable.
2º Represente gráficamente esta variable.
3º Obtenga la media, mediana, moda, variancia, desviación estándar, coeficiente de variación, coeficiente de asimetría, coeficiente de curtosis.
4º Determine el recorrido intercuartílico, interdecílico e intercentílico.
5º Determine el salario por debajo del cual están el 35% de los asalariados.
6º Determine de forma aproximada que porcentaje de asalariados cobran más de una vez y media la el salario medio.
7º Determine el porcentaje de asalariados que cobran salarios comprendidos entre la media menos dos veces la desviación estándar y la media más dos veces la desviación estándar.
8º Obtenga la curva de concentración de Lorenz.
9ª Determine el Indice de Gini.
10ª Responda a todos los apartados anteriores si la unidad monetaria fuera la peseta.

3.- Se les ha preguntado a los 1000 establecimientos de comercio minorista de una determinada ciudad por sus ventas anuales de un cierto producto de alimentación. Los resultados obtenidos son los que refleja la tabla siguiente:

Ventas (Euros)	Establecimientos
Hasta 600	400
De 600 a 1500	225
De 1500 a 3000	175
De 3000 a 6000	120
De 6000 a 9000	75
Más de 9000	5

Además se sabe que el total de ventas para ese producto ascendió a 1953000 euros. Con toda esta información:

a) Según la forma de la distribución, indique y calcule el promedio más adecuado para representar a esta variable.
b) 0btenga la media aritmética de esta variable. Analice su representatividad en términos absolutos y relativos.
c) Determine el porcentaje de establecimientos, de entre los que menos venden, cuyas ventas acumuladas representan la cuarta parte de las ventas totales. Determine también el valor de la variable que deja a su derecha el 10% de los establecimientos que más venden.
d) Si para otra ciudad y esa misma variable se hubieran obtenido unas ventas medias de 3000 euros con una desviación estándar de 6000 € y un índice de Gini de 0,7, realice una análisis comparado de la dispersión y de la concentración de esa variable.
e) Si al siguiente año el precio de ese producto aumenta un 5%, determine cual sería el nuevo volumen medio de ventas así como su variancia.