- **García Ferrando**, M. (1992) *Socioestadística* (Madrid: Alianza) Capítulo 3: "Características de una distribución de frecuencias: tendencia central, dispersión y forma. La distribución normal".

CAPÍTULO 3
CARACTERÍSTICAS DE UNA DISTRIBUCIÓN DE FRECUENCIAS: TENDENCIA CENTRAL, DISPERSIÓN Y FORMA. LA DISTRIBUCIÓN NORMAL

La observación visual de las representaciones gráficas de las distribuciones de frecuencia es, sin duda alguna, un método elemental y aproximado para el análisis de sus propiedades. E1 investigador necesita, a tal fin, disponer de procedimientos de medición más precisos para estudiar las características más sobresalientes de las distribuciones de frecuencias, así como tener un buen conocimiento de los posibles sesgos que puedan introducirse al utilizar tales instrumentos de medición. En el presente capítulo estudiaremos los instrumentos de medida utilizados para caracterizar las distribuciones de frecuencias.

3.1. CARACTERÍSTICAS DE LA DISTRIBUCIÓN UNIVARIABLE

Vamos a aproximarnos a este tema a través de 1a exposición de un ejemplo basado en una investigación real. En un intento por desarrollar una medida fiable y relevante para el estudio de variables sociopsicológicas, Díez Nicolás y Torregrosa (1967, págs. 77 y sigs.) aplicaron la escala de Cantril en la realización de una encuesta sobre «El mundo en el año 2000», tal como es imaginado por la población española. La escala consiste en un *continuum* y se le pide al sujeto que defina, sobre la base de sus propios supuestos, percepciones y valores, los dos extremos de lo «bueno» y lo «malo» o de lo «mejor» y de lo «peor» en relación a un tema concreto.

En el caso concreto del estudio de Díez Nicolás y Torregrosa, el entrevistado sitúa en el extremo superior de la escala sus deseos y esperanzas tal como él mismo las concibe, y cuya realización constituiría «la mejor vida» posible para él. En el otro extremo, el entrevistado expresa sus miedos y preocupaciones, es decir, «lo peor» que podría ocurrirle. Una vez establecidos estos dos puntos extremos, y utilizando e1 *continuum* de 1 a 9, se le preguntó a cada entrevistado dónde creía que estaba situado en la actualidad, dónde creía que estaba situado hace cinco años y dónde creía que se situaría dentro de cinco años.

La aplicación del instrumento de medida a una muestra de 110 personas produjo los siguientes resultados (tabla 1):

TABLA 1

Distribución de las posiciones asignadas por el entrevistado en la escala de Cantril a sí mismo en el momento presente, hace cinco años y dentro de cinco años

Escala de Cantril	*Pasado (%)*	*Presente (%)*	*Futuro (%)*
1	6	—	1
2	7	3	—
3	16	4	2
4	11	16	4
5	25	21	12
6	11	25	16
7	12	18	31
8	6	4	18
9	4	7	14
No sabe, no contesta	2	2	2
TOTAL	(100)	(100)	(100)

FUENTE: J. DÍEZ NICOLÁS y J. R. TORREGROSA, «Aplicación de la Escala de Cantril en España», *REOP*, 1967, pág. 84.

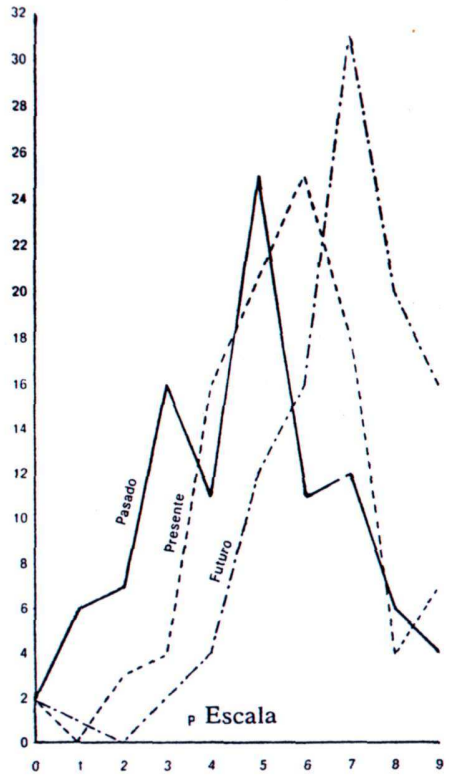
La distribución porcentual de las respuestas pone de manifiesto la existencia de un cierto optimismo al evaluar la población su propia posición en la dimensión temporal. La puntuación asignada tiende a ser mayor a medida que se pasa del pasado al presente y del presente al futuro.

En base a los anteriores datos, y en un estudio sobre la imagen del inundo futuro, ambos autores se sirvieron de dicha investigación exploratoria para formular nuevas hipótesis sobre este tema. Con el fin de visualizar mejor los resultados obtenidos, Díez Nicolás y Torregrosa realizaron la siguiente representación gráfica de la distribución de Frecuencias porcentual contenida en la tabla 1 (ver fig. 1).

Se observa que las tres líneas de grafos son claramente diferentes en una serie de rasgos. En primer lugar, difieren en la *posición* o concentración a lo largo de la escala de puntuaciones. El grafo correspondiente al pasado muestra las puntuaciones más bajas, mientras que el grafo correspondiente al futuro tiene las puntuaciones más altas. En segundo lugar, las tres distribuciones difieren en la relativa *concentración* de las puntuaciones que representan. Así, la distribución correspondiente al pasado está más «apilada» en el centro de la distribución, mientras que las distribuciones correspondientes al presente y al futuro se «apilan» más hacia la derecha y tienen menos casos en la parte izquierda de la

FIGURA 1

Distribución de frecuencias porcentuales de las posiciones asignadas por el individuo en la escala de Cantril a sí mismo en el momento presente, en el pasado y para el futuro



FUENTE: J. DÍEZ NICOLÁS y J. R. TORREGROSA, *op. cit.*, pág. 83.

escala que el grafo del pasado. Además, el grafo del pasado está más disperso que los otros dos, ya que tiene, en general, frecuencias más bajas en las categorías centrales, aunque las tiene más altas para las categorías más bajas. En tercer lugar, la *forma* de las distribuciones también difiere en ciertos aspectos, tales como el número de picos, el grado de asimetría, etc. Estos tres rasgos de las distribuciones se conocen con los nombres de *tendencia central* (o posición), *variación* y *forma*.

En este caso, las diferencias entre las tres distribuciones se han mostrado de la forma que hemos visto en el capítulo anterior, al presentar las técnicas gráficas. En el presente capítulo vamos a ocuparnos de presentar otras formas más compactas para caracterizar las distribuciones de frecuencia que lo que permiten las técnicas gráficas. Y lo haremos a través de la utilización de unas pocas medidas o «números índices» que indican la tendencia central, la variación y la forma de una distribución. De esta manera, la comparación entre diferentes distribuciones se hace más fácil y eficaz, y permite precisar mejor los aspectos en que se asemejan y difieren entre sí las distribuciones de frecuencia.

3.2. LA POSICIÓN DE UNA DISTRIBUCIÓN: MEDIDAS DE TENDENCIA CENTRAL

La posición o «tendencia central» de una distribución se refiere al lugar donde se centra una distribución particular en la escala de valores. Supongamos que tenemos los siguientes cuatro conjuntos de valores referentes a los resultados de unas pruebas en unos grupos de estudiantes. Los seis estudiantes en el grupo c) tienen, en general, puntuaciones más bajas que los de a) o b), mientras que los estudiantes que componen el grupo d) muestran puntuaciones más elevadas:

grupo a)	2	3	3	3	5	5	N=6
grupo b)	2	2	4	5	5	6	N=6
grupo c)	2	2	2	3	4	5	N=6
grupo d)	4	5	6	7	8	8	N=6

Esta comparación resulta cierta a pesar de que algunos estudiantes en c) tienen puntuaciones más altas que en a) y b), y que algunos estudiantes de d) tienen puntuaciones iguales o más bajas que los de a), b) y c). La posición se suele medir a través de una puntuación central o «valor típico» de la distribución, alrededor del cual el resto de los valores tienden a agruparse de una forma determinada. Tres son las medidas de tendencia central más utilizadas, la *moda*, la *mediana* y la *media*, pudiéndose distinguir diferentes tipos de medias, tal como la *media aritmética,* la *media geométrica* y la *media armónica*.

3.2. 1. *Moda*

La moda de una distribución de números es aquel valor que se presenta u ocurre con la mayor frecuencia. Es decir, la moda es el valor más común de la distribución. La moda puede no existir en una distribución determinada o bien puede no ser única. En una representación gráfica, la moda será el rectángulo más alto, en el caso de un histograma, y el pico más alto, en el caso de un polígono.

En el caso del grupo a) anterior, la moda sería el valor 3, mientras que en el caso del grupo b) aparecen dos modas, el 2 y el 5. Las distribuciones que contienen una sola moda se llaman *unimodales*, y las distribuciones que contienen dos modas se denominan *bimodales*. En general, cuando una distribución contiene diversas modas se denomina *multimodal.*

En el caso de datos agrupados, la moda es el punto medio de la clase que contiene la mayor frecuencia de casos. A la clase que contiene la moda se la denomina clase modal. Así, en el ejemplo siguiente, la clase modal será la 4-6 y la moda valdrá 5:

	f_1
De 9 a 11	6
De 7 a 8	10
De 4 a 6	15
De 1 a 3	4
TOTAL	35

ya que la clase 4-6 contiene la mayor frecuencia de casos, 15, y el punto medio entre 4 y 6 es 5.

Si los datos aparecen medidos a nivel nominal, la moda es la categoría a la que corresponde la frecuencia máxima. Así, en el momento de nacer, los niños representan el valor modal, pues nacen más niños que niñas.

Si los datos aparecen medidos a nivel ordinal, la moda es el valor ordinal al que corresponde frecuencia máxima. Así, en la siguiente distribución de frecuencias, que refleja los diferentes grados de acuerdo con un tema determinado, la moda será el valor ordinal «bastante de acuerdo», ya que en él se concentra el mayor número de contestaciones:

	f_1
Muy de acuerdo	15
Bastante de acuerdo	60
Ni poco ni mucho	20
Bastante en desacuerdo	18
Muy en desacuerdo	2
TOTAL	115

La moda tiene, en términos generales, la virtud de ser fácilmente reconocible por simple inspección, por lo que se utiliza como el índice más rápido y directo para determinar la posición de una distribución. Tiene, sin embargo, el inconveniente de no ser necesariamente única -es e1 caso de las distribuciones multimodales- y, además, no es calculable si todos los valores numéricos son diferentes.

3.2.2. *Mediana*

La mediana es el punto o valor numérico que deja por debajo (y por encima) a la mitad de las puntuaciones de una distribución. Así, en la distribución de números siguiente: 5, 6, 7, 8, 9, la mediana es 7, ya que este valor numérico divide exactamente en dos mitades la distribución que tiene un número impar, N=5 de puntuaciones. En general, cuando el número de casos N de la distribución es impar, la mediana se calcula mediante la expresión

$$K=\frac{N+1}{2};$$

de este modo, K nos dará el valor 2 de la posición de la puntuación en la distribución que es la mediana. En el caso anterior, en que

$$K=\frac{5+1}{2}=3,$$

esto es, la mediana será el valor que ocupa la tercera posición; en nuestro caso, el 7.

Si el número de puntuaciones N de la distribución fuera par, como en el siguiente caso: 10, 15, 50, 75, 90, 100, en el que N-6, la mediana sería igual a un valor que se encontrará entre las puntuaciones centrales 50 y 75. En tal caso, el procedimiento habitual de cálculo de la mediana es tomar la media de los dos casos centrales como la media, es decir:

$$M_d = \frac{50+75}{2} = 62\,5$$

En el caso de distribuciones agrupadas en intervalos, la mediana se calcula habitualmente bajo el supuesto de que los casos en el intervalo que contiene la mediana se distribuyen en él homogéneamente. Esto es, que si en un intervalo tenemos cuatro casos, suponemos que cada uno le ellos ocupa la cuarta parte del mismo. La fórmula mediante la que se calcula la mediana con datos agrupados es la siguiente:

$$M_d = L_{md} + \left(\frac{\frac{1}{2} N - acum\, f_{md}}{f_{md}} \right) \cdot W \qquad [3.1$$

en donde L_{md} es el límite inferior del intervalo o categoría que contiene la mediana; N es el número total de casos; $acum_{fmd}$ es la frecuencia acumulada por debajo de la frecuencia del intervalo que contiene la mediana, y W es la amplitud o distancia de la categoría que contiene la mediana.

Se trata de una fórmula similar a la utilizada para calcular los percentiles, dado que, después de todo, la mediana no es otra cosa que el percentil 50. Veamos, a través de un ejemplo, cómo se calcula la mediana en una distribución de datos agrupados. Lo primero que hay que hacer a partir de la distribución de datos dada es la creación de una distribución de frecuencias acumulada, comenzando por la categoría de puntuaciones más bajas:

Puntuaciones	*Frecuencias*	*Frecuencias acumuladas*	*Límites reales*	*Amplitud intervalo*
De 32 a 36	18	88	31,5-36,5	5
De 27 a 31	21	70	26,5-31,5	5
De 22 a 26	26	49	21,5-26,5	5
De 17 a 21	15	23	16,5-21,5	5
De 12 a 16	8	8	11,5-16,5	5

El número de casos que cae por debajo de la mediana será N/2, esto es, 88/2=44. El intervalo que contiene la mediana será aquel cuya frecuencia acumulada está más próxima a 44. En la distribución anterior es el intervalo 22-26 el que contiene la mediana, ya que su frecuencia acumulada, 49, es el número más próximo a 44.

Si la frecuencia acumulada de la categoría que contiene la mediana hubiera sido exactamente igual a N/2, entonces el límite superior del intervalo hubiera sido la mediana. Pero como esto no suele ocurrir habitualmente, como en nuestro ejemplo, se hace preciso recurrir a la anterior fórmula para calcular la mediana. Continuemos, pues, con los cálculos.

La lógica del cálculo es que deseamos localizar un valor, el de la mediana, dentro del intervalo que la contiene, que se encuentra a cierta distancia en el intervalo. La distancia depende de la proporción de frecuencia en el intervalo de la mediana que se necesita añadir a la frecuencia acumulada por debajo del intervalo de la mediana, con el fin de igualar el valor N/2 o el número de casos que caen por debajo de la puntuación de la mediana. Esta proporción se calcula, siguiendo la fórmula, del siguiente modo:

$$\frac{N}{2} - acum\ f_{md} = \frac{88}{2} - 23 = 21.$$

Dado que f_{md} = 26 y L_{md} = 21,5, tal como se observa en el cuadro que contiene las distribuciones, el valor de la mediana será:

$$M_d = 21{,}5 + \frac{21}{26} \cdot 5 = 25{,}5$$

Así, pues, 25,5 será el valor de la puntuación por debajo de la cual queda el 50 por 100 de los casos, esto es, 44 de los 88 casos.

Por todo lo que se ha dicho, queda claro que los valores de una distribución de frecuencias deben tener, como mínimo, un nivel de medición ordinal para que se pueda calcular la mediana, ya que el concepto de la misma implica dirección (puntuaciones por arriba y por debajo de la mediana). Ahora bien, la mediana es un índice de posición que no presupone conocimiento de la distancia, excepto para el caso de la amplitud del intervalo en el que cae la mediana cuando se tienen datos agrupados. Esto quiere decir que si se utiliza con datos medidos a nivel de intervalo, se pierde algo de información, al igual que ocurre si utilizamos la moda con tal tipo de datos. En cierto modo, esto constituye una ventaja de la mediana, ya que es poco influida por la existencia de valores extremos altos y erráticos, ya que es simplemente el punto que divide a todos los casos en dos mitades. En el caso de datos agrupados, la mediana se puede calcular aunque la categoría o intervalo máximo no tenga límite superior ni la categoría o intervalo mínimo lo tenga superior, siempre que la mediana no caiga en tales categorías y extremos, lo que, por otro lado, no es corriente.

La mediana es fundamento de diversas técnicas estadísticas, aunque el número y utilización de éstas es notablemente menor que el de las técnicas basadas en la media aritmética, que va a ser estudiada a continuación.

3.2.3. *Media aritmética*

La media común o media aritmética es, simplemente, la suma de todas las puntuaciones de una distribución dividida por el número de casos. Así, dados n valores, X_1, X_2,..., X_n, su media aritmética, X, viene definida por:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad [3.2]$$

De una manera más simplificada, se puede escribir la media, prescindiendo de los subíndices en el sumatorio, mediante la expresión:

$$\bar{X} = \frac{\Sigma X}{n}$$

sobreentendiéndose que ∑X, sin ningún subíndice, indica la sumación de dos los valores.

La media aritmética posee algunas características muy interesantes, que la hacen muy útil y la medida más ampliamente utilizada de tendencia central.

Para comenzar, la media aritmética es otro buen ejemplo del uso estadístico de las razones como una forma válida de realizar comparaciones. La suma total de las puntuaciones se «estandariza», por decirlo de alguna forma, en términos de las puntuaciones que se incluyen en la suma. Esto permite comparar las medias de grupos de diferente tamaño, mientras que la comparación directa de las correspondientes distribuciones sería errónea. Algunas veces, no obstante, el número de puntuaciones que contribuyen a la suma no es la única fuente productora de diferencias al realizar comparaciones de tendencia central entre grupos. En el cálculo de la suma, cada puntuación contribuye en una forma o cantidad diferentes, dependiendo de su valor numérico. Naturalmente, las puntuaciones elevadas contribuyen más a la suma que las puntuaciones bajas, lo que significa que los valores extremos elevados tienen una influencia mayor en el cálculo de la media que las puntuaciones intermedias más bajas. Se puede decir que la media es «atraída» por los valores extremos altos en una distribución. Así, supongamos que tenemos la siguiente distribución: a) 2, 2, 4, 6, 8, 14, 20, cuya media X=8,0. Pues bien, basta que el valor numérico del extremo pase de 20 a 30 -quedando entonces la distribución como b) 2, 2, 4, 6, 8, 14, 30- para que la media cambie significativamente su valor, X=9,4, es decir, 1,4 unidades superior a la anterior.

Por esta razón, se ha comparado a la media como el punto de apoyo o fulcro de un tablero ideal e imaginario en el que quedan situados, a derecha e izquierda del fulcro, los valores que están situados por encima o por debajo de la media. En otras palabras, se puede describir a la media como el «centro de gravedad» de la distribución de frecuencias (Amón, 1973, pág. 50).

En algunos casos, interesa asociar a los números, X_1, X_2,..., Xn, ciertos factores o pesos, W_1, W_2,..., W_n, que dependen de la significación o importancia de cada uno de los números. En tal caso, la media se calcula mediante la expresión:

$$\bar{X} = \frac{W_1X_1 + W_2X_2 + \ldots + W_nX_n}{W_1 + W_2 + \ldots + W_n} = \frac{\Sigma WX}{\Sigma W} \qquad [3.3]$$

A este tipo de media se la denomina *media aritmética ponderada*. Su uso viene aconsejado cuando se pretende calcular la media en una distribución cuyos valores tienen diferente significado o importancia de cara al resultado final. Supongamos que los resultados de un examen final dependen de tres exámenes parciales que se valoran de forma distinta; por ejemplo, el último de ellos es tres veces más importante que los dos primeros. Si las notas obtenidas en el primer, segundo y tercer examen por un alumno concreto han sido 6, 5 y 7, respectivamente, la nota media final o media ponderada será:

$$\bar{X} = \frac{(1)(6)+(1)(5)+(3)(7)}{1+1+3} = \frac{32}{5} = 6{,}4$$

Veamos ahora un ejemplo real de utilización de la media ponderada. En un estudio sobre la conciencia regional de los españoles se encontró la siguiente distribución porcentual de autoubicación, en un espacio político izquierda-derecha:

	% del total	% del total menos los % de NS/NC
Izquierda: 1	2	3
2	3	4
3	6	8
4	7	9
5	24	30
6	14	18
7	6	8
8	7	9
9	4	5
Derecha: 10	5	6
No sabe	14	
No contesta	7	100
	100	
	(6.342)	

FUENTE: J. JIMÉNEZ BLANCO et al., *La conciencia regional de España*, Madrid, C.I.S., 1977. Elaboración propia.

Con el fin de calcular la media nacional de la autoubicación en la escala izquierda-derecha, se hace preciso considerar el porcentaje de población que se autoubica en cada una de las casillas de la escala. Ahora bien, como en la distribución original existe un 21 por 100 de entrevistados que no han respondido -14 por 100 por «no sabe» y 7 por 100 por «no contesta»-, es necesario volver a calcular la distribución porcentual en base a los que sí se han autoubicado, distribución que aparece en la columna de la derecha de la tabla. Con estos datos ya se puede calcular la media

ponderada, que nos dará el valor de la posición media de la población española en dicha escala:

$$\bar{X}=\frac{1\cdot 3+2\cdot 4+3\cdot 8+4\cdot 9+5\cdot 30+6\cdot 18+7\cdot 8+8\cdot 9+9\cdot 5+10\cdot 6}{100}=5{,}64$$

Así, pues, si consideramos que el centro político se encuentra entre las casillas 5 y 6, se puede afirmar que la media nacional, con un valor de 5,64, es claramente centrista desde el punto de vista político.

Otras propiedades interesantes de la media aritmética son las siguientes:

a) La suma algebraica de las desviaciones de un conjunto de números con respecto a su media aritmética es igual a cero. Es decir, dada una media aritmética X=K, la suma de las diferencias de las n puntuaciones X_1, X_2,..., X_n, respecto a K, vale 0. En efecto, se tiene que:

$$\Sigma (X_i-\bar{X})=\Sigma X_i-n\bar{X}=\Sigma X_i-n\frac{\Sigma X_i}{n}=\Sigma X_i-\Sigma X_i=0$$

b) Si la suma de los cuadrados de las desviaciones de un conjunto n de números X_1, X_2,..., X_n, respecto a K es mínima, entonces K= X, ya que si K fuera distinto de la media aritmética, la suma de las diferencias al cuadrado no podría ser mínima, tal como se ha visto anteriormente.

c) Si n_1 números tienen de media m_1; n_1 números tienen de media m_2,...; n_i números tienen de media m_i, entonces la media de todos los números es:

$$\bar{X}=\frac{n_1m_1+n_2m_2+\ldots+n_im_i}{n_1+n_2+\ldots+n_i} \qquad [3.4]$$

es decir, se trata de una media ponderada de todas las medias posibles del conjunto de números.

d) Si la media $Y_1 = AX_1 + B$, la media de $Y_2 = AX_2 + B$,..., y la media de $Y_n = AX_n + B$, siendo A y B dos constantes arbitrarias, entonces la media de todas las Y_i es $Y = AX + B$, ya que, por definición (siendo i = 1, 2,... n):

$$\bar{Y}=\frac{\Sigma Y_i}{n}=\frac{\Sigma (AX_i+B)}{n}=\frac{A\Sigma X_i+nB}{n}=A\frac{\Sigma X_i}{n}+\frac{nB}{n}=A\bar{X}+B \qquad [3.5]$$

Cuando los datos se presentan agrupados mediante una distribución de frecuencias, todos los valores caen dentro de unos intervalos de clase que, a efectos de cálculo, se consideran coincidentes con los puntos medios de cada intervalo. Para el caso en que todos los intervalos sean de idéntica amplitud, y siendo X, el punto medio de cada

intervalo y *f* la frecuencia, la *media aritmética de datos agrupados* se calcula mediante la expresión:

$$\bar{X} = \frac{fX_i}{N} \qquad [3.6]$$

Veamos a través de un ejemplo la utilización práctica de dicha fórmula. A partir de la distribución de frecuencias dadas se crea una columna de puntos medios:

Puntuaciones	X_i	f
De 22 a 26	24	18
De 17 a 21	19	21
De 12 a 16	14	26
De 7 a 11	9	15
De 2 a 6	4	8
		$N=88$

Con estos datos ya estamos en condiciones de aplicar la fórmula [3.6]

$$\bar{X} = \frac{18\cdot 24+21\cdot 14+26\cdot 14+15\cdot 9+8\cdot 4}{88} = \frac{1.362}{88} = 15{,}5$$

3.2.4. *Tipos especiales de medias*

Existen otras medidas de tendencia central que son apropiadas para situaciones especiales que, sin embargo, son más corrientes en las ciencias físicas que en las ciencias sociales. De todos modos, algunas veces pueden ser utilizadas por los sociólogos, por lo que expondremos aquí su definición y forma de cálculo.

La *media geométrica* de una serie N de números X_1, X_2,..., X_n, es la raíz n-ésima del producto de los números:

$$\text{Media geométrica} = \sqrt[n]{(X_1)(X_2)\dots(X_n)} \qquad [3.7]$$

En la práctica, la media geométrica se calcula mediante logaritmos *. Su uso es apropiado cuando hace falta calcular la razón media de varias razones, como ocurre en algunas técnicas de construcción de escalas de actitudes, o cuando se desea calcular el porcentaje medio de cambio de alguna característica variable. Obsérvese que la media geométrica se calcula de forma parecida a la media aritmética, cambiando tan sólo los signos de suma y división de ésta por los signos de multiplicación y radicación en aquélla.

La *media armónica* de una serie N de números X_1, X_2,..., X_n, es el número recíproco de la media aritmética de los recíprocos de los números:

$$\text{Media armónica} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i}} = \frac{N}{\sum_{i=1}^{N}\frac{1}{X_i}} \qquad [3.8]$$

El uso de la media armónica puede resultar de utilidad en problemas que tengan que ver con cambios en el tiempo, distancias, etc.

La *media cuadrática* es un valor tal que su cuadrado es igual a la media aritmética de los cuadrados de los números:

$$\text{Media cuadrática} = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{N} \qquad [3.9]$$

El uso de la media cuadrática tiene interés en el cálculo de la varianza, de la que nos ocuparemos más adelante.

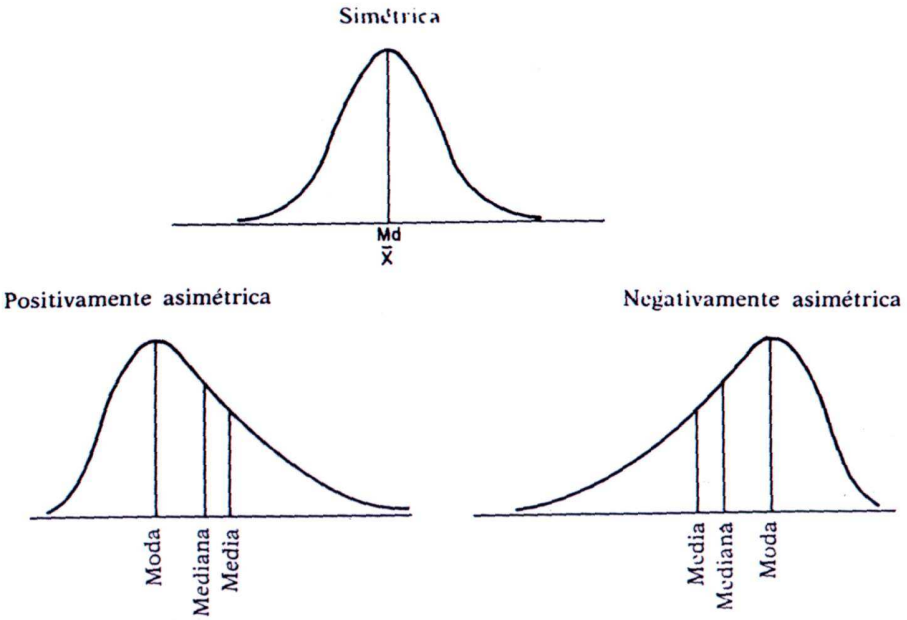
3.2.5. *Relación y comparación entre los índices de tendencia central*

Hemos visto anteriormente que la media utiliza más información que la mediana, en el sentido de que todas las puntuaciones entran en el cálculo de la media, mientras que el cálculo de la mediana tan sólo implica la puntuación del caso medio. De ahí que la media quede afectada por cambios en los valores extremos, cosa que no ocurre en el caso de la mediana.

Esta importante diferencia entre la media y la mediana permite, en muchos casos, poder tomar una decisión sobre cuál de ellas resulta más apropiada. En un principio, suele resultar más apropiado para el investigador el poder hacer uso de toda la información que se contiene en la distribución de frecuencias por lo que, desde este punto de vista, resulta más ventajoso el empleo de la media que el de la mediana. Además, la media es una medida más estable que la mediana, en el sentido de que varía menos de una muestra a otra. Este es un tema que estudiaremos con mayor atención cuando nos ocupemos de la estadística inductiva. Baste decir aquí que cuando se trabaja con una muestra proveniente de una población, lo que le interesa principalmente al investigador es poder generalizar los resultados de la muestra a la población. Se sabe que si se hubiera tomado otra muestra los resultados no serían ya los mismos. Sólo si se pudiera tomar una serie de muestras podríamos saber cuánto difieren entre sí las medias de las diferentes muestras. Lo que afirmamos ahora es que la media diferirá menos de una muestra a otra de la misma población que lo hará la mediana. En conclusión, pues, el uso de la media suele ser preferible al de la mediana como medida de tendencia central.

Ahora bien, si la distribución es muy asimétrica, se corre el peligro de que un valor extremo muy alto altere profundamente el valor de la media, distorsionando su sentido. En tal caso, el uso de la mediana está más recomendado, ya que ofrecerá una mejor descripción del carácter de la distribución. Las posiciones relativas de la media y de la mediana dependen, pues, del tipo de simetría-asimetría de la distribución. En las

distribuciones perfectamente simétricas, la media y la mediana coinciden, mientras que en las distribuciones asimétricas las posiciones relativas de ambos índices varían según el sesgo de la asimetría, tal como se observa en las siguientes figuras:



Pero no sólo varía la posición relativa de la moda y de la mediana según la forma y grado de asimetría, sino que también lo hace la posición de la moda, tal como se puede observar en las figuras anteriores. Se puede demostrar que para curvas de frecuencias unimodales que sean moderadamente sesgadas se cumple la siguiente relación empírica: Media-Moda=3 (Media-Mediana).

Para terminar esta exposición sobre las medidas de tendencia central destaquemos, una vez más, la importancia estadística de la media aritmética, por ser parte integrante de la lógica seguida en la creación de otros procedimientos estadísticos, tales como la varianza y la desviación típica, la correlación y regresión y el análisis factorial, que tendremos ocasión de estudiar en capítulos subsiguientes.

3.3. VARIACIÓN O DISPERSIÓN DE UNA DISTRIBUCIÓN

Si realizáramos un estudio comparativo sobre el origen social de los estudiantes universitarios españoles en 1980 y en 1950 y midiéramos el origen social de los estudiantes por medio de una escala del prestigio ocupacional de los padres, el proceso de masificación y la relativa democratización de la universidad española experimentado en el período 1950-1980, se reflejarían en una mayor dispersión y variación de los valores de la escala de prestigio ocupacional de los padres, por lo que se refiere a los estudiantes matriculados en 1980, en relación a los que estaban matriculados en 1950. Y ello como consecuencia de la afluencia a la universidad en mayor proporción de

estudiantes de clase social media y obrera, lo que se ha traducido en una ampliación de los estratos sociales que envían alumnos a la universidad.

Para medir ese rasgo diferenciador de las distribuciones de frecuencias correspondientes a los dos extremos del período considerado, hace falta recurrir a medidas que den cuenta del grado de dispersión o variación de las puntuaciones. Así como las medidas de tendencia central o posición indican dónde se sitúa un grupo de puntuaciones, los índices de variabilidad o dispersión indican si esas puntuaciones son muy parecidas o muy distintas entre sí. Por ejemplo, las tres siguientes distribuciones:

a)	51	52	53	54	55	*N=5*
b)	52	53	53	53	54	*N=5*
c)	47	50	53	56	59	*N=5*

tienen idéntica media y mediana, 53, y, sin embargo, los tres grupos difieren entre sí en el grado de agrupamiento-dispersión de sus puntuaciones alrededor del valor medio. El grupo c) está claramente más disperso que los grupos a) y b).

Existe una diversidad de formas de cálculo para la medición de la variabilidad en un grupo de puntuaciones, distinguiéndose las diferentes formas de cálculo según se trate de datos nominales, ordinales o de intervalo. Frecuentemente, la variación en las distribuciones ordinales se mide a través de las mismas técnicas utilizadas con datos de intervalo, a pesar de que la distancia entre puntuaciones no está definida con los datos ordinales. Vamos a comenzar el estudio de las técnicas de medición de la variación o dispersión con las utilizadas en los datos de intervalo. En este caso se siguen dos procedimientos, según se considere el recorrido o amplitud de la escala en la que se distribuyen las puntuaciones, o bien se describa la variación por medio de las diferencias que se producen entre todas las puntuaciones y un índice de tendencia central. Veamos a continuación los primeros.

3.3.1. *Recorrido*

El recorrido o rango de un conjunto de números es, simplemente, la diferencia entre el mayor y el menor de todos ellos. Así, si disponemos de la distribución de los sueldos que perciben los empleados de una empresa de forma tal que el sueldo más elevado sea 150.000 pesetas y el sueldo más bajo 45.000 pesetas el recorrido de los sueldos de dicha empresa será: 150.000-45.000=105.000 pesetas.

La desventaja de esta medida es que sólo depende de los valores extremos de una distribución y no tiene en cuenta los valores intermedios. Si se trata de dos valores atípicos, por ejemplo, en el caso anterior, los sueldos del gerente y del aprendiz, la medida del recorrido no nos dice nada acerca de los valores de los sueldos de los empleados de la fábrica. Por ello se utilizan otras medidas que tengan en cuenta un mayor volumen de la información que contienen las distribuciones.

El *recorrido intercuartílico*, o diferencia entre los cuartiles tercero y primero, mejora la medida del recorrido o rango ordinario, porque, al tratarse de cuartiles, son más sensibles a la propia concentración de los datos. Recuérdese que el primer cuartil Q_1 es el punto de la escala debajo del cual queda el 25 por 100 de los casos, mientras que debajo del tercer cuartil Q_3 queda el 75 por 100 de los casos. Por tanto, entre el recorrido

intercuartílico Q_3-Q_1 queda el 50 por 100 de los casos. Alcaide (1976, pág. 143) calcula el recorrido intercuartílico de la distribución por edades de los españoles censados en 1970 en 35,74 años, ya que el primer cuartil se encuentra en la edad 13,44 y el tercer cuartil en la edad 49,18. En consecuencia:

$Q_3 - Q_1 = 49{,}18 - 13{,}44 = 35{,}75$ años

y, tal como se ha dicho anteriormente, en este recorrido de edades se encuentra el 50 por 100 de la población española.

A veces se utiliza como medida de dispersión el *recorrido semiintercuartílico* o *desviación cuartílica*, que viene definido por la mitad del recorrido intercuartílico; esto es:

$$\text{Recorrido semiintercuartílico} = \frac{Q_3 - Q_1}{2} \qquad [3.10]$$

El recorrido intercuartílico tiene la ventaja sobre el recorrido ordinario, tal como se ha dicho antes, de evitar el uso exclusivo de las dos puntuaciones extremas y de estar menos sujeto, por tanto, a la variación errática de tales valores. También se pueden calcular las distancias entre otros dos puntos significativos. Así, por ejemplo, el recorrido entre percentiles 10-90 de una distribución de frecuencias viene definido por la diferencia entre el percentil nonagésimo P_{90} y el percentil décimo P_{10}. Tiene parecidas ventajas que el recorrido intercuartílico.

3.3.2. *Desviación media*

La *desviación media* o *promedio de desviación* es otra medida de dispersión que viene dada por la media aritmética de los valores absolutos de las desviaciones observadas a un determinado valor medio. Así, dada tina serie N de números X_1, X_2,..., N_n la desviación media DM viene definida por:

$$DM = \frac{\Sigma f_i |X_i - \bar{X}|}{N}$$

donde X es la media aritmética de los números dados y $|X_i - X|$ es el valor absoluto de las desviaciones de los diferentes valores de X al valor medio X. (Recuérdese que el valor *absoluto* de un número es el mismo número sin asociarle signo alguno, y se indica por dos barras verticales a ambos lados del número. Así, |-51=5, 1+3|=3, |7|=7.)

Para calcular la desviación media de los números o conjunto de observaciones siguiente: (2, 4, 6, 8, 10), se calcula en primer lugar su media aritmética:

$$X = \frac{2+4+6+8+10}{5} = 6$$

y a continuación se calcula la desviación media respecto a la media aritmética:

$$MD=\frac{|2-6|+|4-6|+|6-6|+|8-6|+|10-6|}{5}=$$

$$=\frac{|-4|+|-2|+|0|+|2|+|4|}{5}=\frac{4+2+0+2+4}{5}=2{,}4$$

Si los números X_1, X_2,..., N_n se presentan con frecuencias f_1, f_2,..., f_k, respectivamente, la desviación media puede escribirse como:

$$DM=\frac{\Sigma f_i|X_i-\bar{X}|}{N}=\frac{\Sigma f|X-\bar{X}|}{N}$$

donde $N = \Sigma f_i = \Sigma f$. Esta expresión es útil cuando se dispone de datos agrupados en donde las diferentes X_i representan los valores medios de clase y las f_i, las correspondientes frecuencias de clase.

En general, se puede afirmar que cuanto mayor sea el valor de la desviación media, mayor será la variación entre las diferentes puntuaciones. Aunque la desviación media se calcula e interpreta fácilmente, existen otras medidas de dispersión que son más preferidas, porque intervienen en la elaboración de otras áreas de la estadística. La varianza es la medida de dispersión más ampliamente utilizada.

3.3.3. *Desviación típica y varianza*

La *varianza* y la desviación típica son medidas similares a la desviación media, en el sentido de que se basan en las diferencias existentes entre la media aritmética y cada puntuación, pero se diferencian de ella en que, en lugar de tomar el valor absoluto de tales desviaciones, se utiliza el cuadrado de las mismas. De esta forma, se logra una medida de dispersión para datos de intervalo que tiene un amplio campo de aplicabilidad en la estadística, por estar relacionado con otros temas estadísticos.

La *varianza* es simplemente el valor medio del cuadrado de las desviaciones de las puntuaciones a la media aritmética, mientras que la desviación típica (en inglés, *standard deviation*) es la raíz cuadrada de la varianza:

$$s^2=\frac{\Sigma(X_i-\bar{X})^2}{N} \qquad [3.12]$$

$$s=\sqrt{s^2}=\sqrt{\frac{\Sigma(X_i-\bar{X})^2}{N}} \qquad [3.13]$$

Nótese una cuestión de símbolos. Cuando se opera con datos muestrales, los símbolos estadísticos con los que se representa la varianza y la desviación típica son los que aparecen en las fórmulas [3.12) y [3.13], esto es S^2 y S; mientras que si los datos hacen referencia directamente a la población general, los parámetros que simbolizan la

varianza y la desviación típica se representan mediante el símbolo σ, que es la letra griega sigma minúscula En tal caso, la varianza será σ^2 y la desviación típica será σ,

Veamos un ejemplo concreto de cálculo. Dado el conjunto de números (2, 2, 4, 6, 8, 14, 20), de media X- 8, el cálculo de la varianza y de la desviación típica requerirá el cálculo previo de las diferencias de cada número respecto a la media y ulterior aplicación de las fórmulas [3.12] y [3.13], del siguiente modo:

X_i		*Diferencias* $X_i - \bar{X}$	*Diferencias al cuadrado* $(X_i - \bar{X})^2$
2		$2-8=-6$	36
2		$2-8=-6$	36
4		$4-8=-4$	16
6	$\bar{X}=8{,}0$	$6-8=-2$	4
8	$N=7$	$8-8=0$	0
14		$14-8=6$	36
20		$20-8=12$	144
56		0	272

$$s^2=\frac{\Sigma(X_i-\bar{X})^2}{N}=\frac{272}{7}=38{,}9$$

$$s=\sqrt{s^2}=\sqrt{38{,}9}=6{,}2$$

El significado intuitivo de una desviación típica de 6,2 no se hará evidente hasta que estudiemos, más adelante, las áreas que quedan por debajo de la curva normal. Por el momento, aceptemos el valor de la desviación típica como un número abstracto, que es tanto más grande cuanto más elevada sea la dispersión de las puntuaciones alrededor de la media aritmética.

Habitualmente, las fórmulas [3.12] y [3.13] no se utilizan a efectos de cálculo porque requieren el cálculo adicional de la media aritmética y de la desviación de cada puntuación a la media -lo que siempre puede introducir una nueva fuente de error-. Las siguientes fórmulas son de uso más práctico, distinguiéndose entre distribuciones de frecuencias que presentan sus datos agrupados de aquellas otras que no los presentan.

Datos no agrupados:

$$s^2=\frac{\Sigma X_i^2-(\Sigma X_i)^2/N}{N} \qquad [3.14]$$

en donde ΣX_i^2 es la suma de las puntuaciones al cuadrado y $(\Sigma X_i)^2$ es el cuadrado de la suma de las puntuaciones. Naturalmente, la desviación típica será la raíz cuadrada de la varianza.

En el ejemplo anterior, $\Sigma X_i^2 = 720$, $(\Sigma X_i)^2 / N = 445$:

$$s^2 = \frac{720 - 445}{7} = 38{,}9$$

Datos agrupados:

La fórmula de la varianza para datos agrupados es similar a la fórmula anterior; sólo que en lugar de las puntuaciones originales se utilizan los puntos medios de la clase y las correspondientes frecuencias. En tal caso, la fórmula para la varianza es como sigue:

$$s^2 = \frac{\Sigma f_i X_i^2 - (\Sigma f_i X_i)^2 / N}{N} \qquad [3.15]$$

en donde $\Sigma f_i X_i^2$ es el sumatorio de los productos de las frecuencias por el cuadrado de los correspondientes puntos medios para todas las clases o categorías, y $(\Sigma fi Xi)^2$ es la suma al cuadrado de los productos de las frecuencias por os correspondientes puntos medios.

Otras propiedades de la desviación típica son las siguientes. Para el caso en que todos los valores de la distribución fueran iguales, las desviaciones de todos los valores alrededor de la media valen cero, y éste será también el valor de la desviación típica. Además, se observa fácilmente que los valores extremos en relación a la media tienen un gran peso en el calculo d la desviación típica, ya que son elevados al cuadrado. Así, en el ejemplo numérico anterior, la puntuación 20 tiene una gran influencia en la determinación de s, ya que, al elevar al cuadrado su diferencia con la media, se convierte en 144, que representa más de la mitad del valor de la suma de todas las diferencias al cuadrado, que en 272. Vemos, pues, que los valores extremos tiene un gran influencia en el valor de s, por lo que, tal como señala Blalock (1960, pág. 8), hay que moderar el entusiasmo inicial con la desviación típica como la mejor medida de una dispersión. Se ahí que el propio Blalock surgiera que cuando una distribución tenga unos pocos casos extremos conviene mas utilizar la mediana o a desviación intercuartílica en lugar de s, como medidas mas apropiadas de dispersión.

Para el caso de datos agrupados existen formulas mas complejas de calculo que las [3.12] y [3.13]. Sin embargo, nos abstenemos de reproducirlas aquí porque en la práctica cada vez se utilizan menos los cálculos manuales, toda vez que el uso masivo de pequeñas, medianas y grandes calculadoras exime cada vez mas al investigador de realizar fatigosos cálculos manuales, sujetos a un margen de error mas grande que el que permiten las calculadores automáticas. Los programas estándar de análisis de datos sociológicos, sobre todo de los provenientes de encuestas, calculan ya, como parte de sus rutinas, la media y la desviación típica, como medidas de dispersión de las distribuciones de frecuencias. Una forma típica de salida de resultados en un análisis de datos de encuesta mediante ordenador es la que se reproduce en la tabla 2, en la que aparecen las calificaciones que a una población, diferenciada según su nivel de religiosidad, le merece una serie de delitos.

La interpretación de los resultados que se contienen en la tabla 2 no es tarea específica de este texto. Baste señalar, sin embargo, que las diferencias mas claras entre

los diferentes grupos de población, diferenciados por su nivel de religiosidad, se producen al evaluar el homicidio y el aborto, mientras que para el caso de la violación y del asesinato premeditado las medias entre los diferentes grupos son análogas y las desviaciones típicas muy bajas. No ocurre así en el caso, sobre todo, del aborto, para el que se observa una actitud claramente más condenatoria entre la población católico-practicante, X = 3,5 y s = 2, que entre la población no creyente, X = 5,2 y s = 2,4. Este es un buen ejemplo de cómo unos valores profundos, como son los religiosos, determinan unas opiniones concretas, en este caso la calificación de unos delitos, y de cómo tales diferencias se hacen estadísticamente evidentes mediante el uso de dos medidas de dispersión o variación.

TABLA 2

Calificación en una escala del 1 al 9 de la gravedad de unos delitos

	TOTAL	*Homicidio por conducir embriagado*			*Violación*			*Asesinato premeditado*			*Aborto a los cinco meses*		
		$\bar{X}$	s_x	%	$\bar{X}$	s_x	%	$\bar{X}$	s_x	%	$\bar{X}$	s_x	%
Religiosidad:													
Católico practicante ...	(9.508)	4,1	1,7	84	3,2	1,6	87	1,6	1,1	89	3,5	2	85
Católico no practicante	(4.880)	4	1,8	89	3,1	1,6	89	1,5	1,1	92	4,1	2,3	84
Creyente otra religión	(98)	4,2	2,3	84	3,1	2	81	1,6	1	83	3,8	2,5	69
No creyente	(392)	3,6	1,6	89	3	1,7	89	1,5	1	95	5,2	2,4	67
Indiferente	(805)	3,8	1,9	88	3,2	1,6	87	1,6	1,4	92	4,6	2,5	74

FUENTE: «Encuesta de victimización», *Revista Española de Investigaciones Sociológicas*, 4, 1978, pág. 245. La escala va del 1, más grave, al 9, menos grave.

Algunas veces puede resultar deseable comparar diversos grupos en relación a su relativa homogeneidad cuando los grupos tienen medias diferentes, pero puede motivar cierta confusión la comparación de las magnitudes absolutas de las desviaciones típicas. Por eso resulta aconsejable utilizar como elemento de comparación la desviación típica en relación a la media. En tal caso, se puede obtener una medida de la variabilidad relativa dividiendo la desviación típica por la media, lo que se denomina *coeficiente de variabilidad V*. Entonces:

$$V = \frac{s}{\bar{X}} \qquad [3.16]$$

Veamos las ventajas del coeficiente de variabilidad sobre la desviación típica mediante la continuación del ejemplo anterior. En relación a la calificación del aborto, los católicos practicantes tienen una media de 3,5 y una desviación de 2, mientras que los no creyentes ofrecen una media de 5,2 y una desviación de 2,4. El coeficiente de variabilidad de ambos grupos será, por tanto, 2/3,5=0,57 y 2,4/5,4=0,44, lo que da una

diferencia más pequeña que la existente entre ambas desviaciones típicas. El coeficiente de variabilidad, llamado también de Pearson, se suele multiplicar por 100 con el fin de ofrecer su valor porcentual. En el ejemplo anterior, la desviación típica del grupo de católicos es el 57 por 100 de la media aritmética, valor superior al 44 por 100 de la media aritmética que vale la desviación típica entre los no creyentes. Vistos así los resultados, la comparación de ambos grupos es más clara que si se hubieran utilizado exclusivamente las desviaciones típicas.

3.3.3.1. *Puntuaciones normalizadas y referencias tipificadas*

En el capítulo anterior vimos los diferentes tipos de comparaciones que se podían realizar. Buena parte de los procedimientos estadísticos que venimos exponiendo en el presente capítulo tratan de facilitar la comparación grupo a grupo o la comparación grupo con tipos estándar. También se puede hacer uso de algunos de los estadísticos estudiados hasta ahora para indicar la relativa posición de un individuo en su grupo. Una de estas formas puede ser el cálculo del rango de percentil de un individuo, esto es, el porcentaje de todas las puntuaciones que son iguales o menores que dicha puntuación. Otra forma de comparar un individuo con un grupo es la creación de *puntuaciones normalizadas* o *típicas*, que se suelen designar mediante la letra minúscula latina *z*. Una puntuación normalizada o típica es simplemente el número de unidades de desviación típica que un individuo queda por encima (o por debajo) de la media de su grupo:

$$z = \frac{(X_i - \bar{X})}{s} \qquad [3.17]$$

También se suelen referir las puntuaciones normalizadas como *variables normalizadas* o *típicas*. En todo caso, y como se ve a través de la fórmula [3.17], en la puntuación normalizada se elimina el efecto de la media (por sustracción) y se expresa la diferencia en unidades de desviación típica, al dividir por ella. Por esta razón, las cantidades de las puntuaciones normalizadas son adimensionales, esto es, son independientes de las unidades empleadas.

En general, cuando las desviaciones de la media vienen dadas en unidades de desviación típica, se dice que están expresadas en *unidades tipificadas* o *referencias tipificadas*. Son de gran valor en el manejo de comparaciones entre distribuciones. Varias son las propiedades de las puntuaciones *z* dignas de interés. La media de dichas puntuaciones es cero, y su desviación típica vale la unidad. Otra propiedad interesante de las puntuaciones *z*, que se utilizará, más adelante, cuando estudiemos el coeficiente de correlación, es que la suma de los cuadrados de las puntuaciones *z* es igual al número N de casos; esto es, que $\Sigma z^2 = N$.

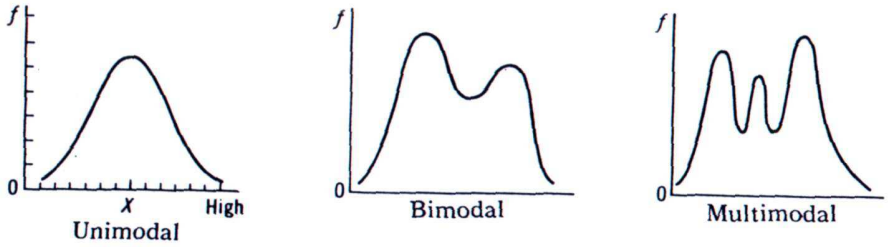
3.4. FORMA DE UNA DISTRIBUCION

El rasgo de una distribución más directamente aparente a partir de un histograma o de un polígono es la forma global de dicha distribución. En general, una distribución de

frecuencias queda bastante bien caracterizada cuando conocemos de ella algún índice de tendencia central y de variabilidad, pero quedará todavía mejor caracterizada si conocemos su grado de simetría-asimetría y su apuntamiento. Veamos a continuación algunas características descriptivas de la forma, de una distribución, y algunos de los índices desarrollados para medir dicha forma.

3.4.1. *Características de la forma de una distribución: Asimetría y apuntamiento*

Una primera característica de la forma de una distribución que, a simple vista, se puede tomar en consideración de un histograma o polígono de frecuencias es el número de picos o puntas (modas) que tiene la distribución. Si la distribución tiene sólo una punta o moda se llamará *unimodal*, y si tiene dos puntos altos se denominará *bimodal*. Obsérvese que la determinación del número de puntas o picos depende, en buena medida, del criterio del investigador en su asignación de importancia a las diferencias en la frecuencia de las categorías. En una distribución *multimodal*, en la que las puntas tengan diferentes alturas -es decir, representan diferentes frecuencias-, corresponde al investigador decidir cuántas modas considera relevantes. En los siguientes gráficos hemos representado algunas de las formas que pueden tomar las distribuciones de frecuencia desde el punto de vista de las puntas o picos que presentan:

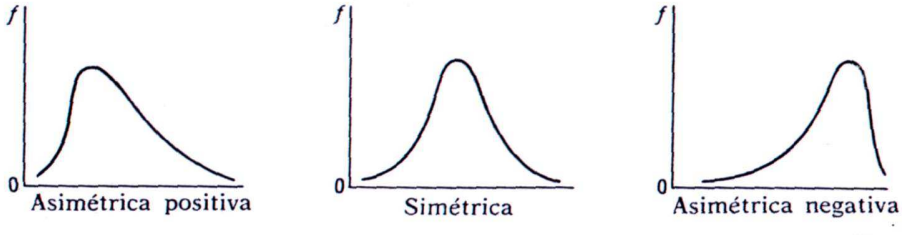


Una segunda característica de la forma de una distribución viene dada por su grado de simetría. La idea general de simetría es bastante sencilla. Sabemos que la mediana divide el histograma en dos áreas de la misma superficie. Pues bien, se dice que la distribución de frecuencias es simétrica cuando una de las áreas es imagen de la otra. Nótese que si un área es imagen de la otra ambas tienen la misma superficie, pero lo contrario no es necesariamente cierto. Es decir, ambas áreas pueden tener la misma superficie pero no representar imágenes recíprocas.

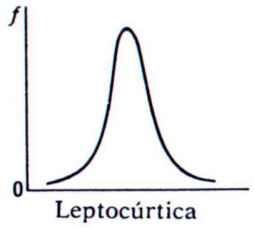
Cuando la curva es simétrica, la mediana coincide con la media. Si, además, la distribución de frecuencias es unimodal, la moda coincide igualmente con la media y la mediana.

Se dice que la simetría es positiva si existen muchas puntuaciones bajas y poco altas, mientras que la simetría es negativa si sucede lo contrario. Si la distribución es asimétrica y unimodal, la mediana y la moda no coinciden. Si la asimetría es negativa, el orden es de izquierda a derecha; es decir, primero está la media, después la mediana y, por último, la moda. Si la asimetría es positiva, el orden es el contrario; esto es, moda,

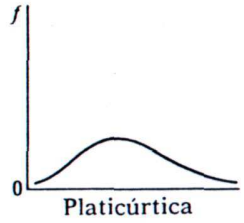
mediana y media. En los siguientes gráficos se representan curvas simétricas y asimétricas:



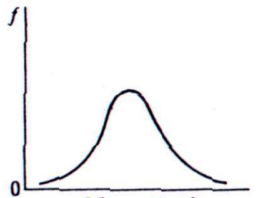
Otro rasgo importante de la forma de una distribución se refiere al grado de apilamiento de los casos alrededor de un punto en la distribución. La *curtosis* hace referencia precisamente al grado de apuntamiento de una distribución. Para el caso de una distribución unimodal y simétrica, la forma *leptocúrtica* aparece cuando presenta un apuntamiento relativo alto, es decir, cuando se tiene una distribución de frecuencias altamente concentrada, como en la figura siguiente:



Si la distribución de las frecuencias es más uniforme, la forma de la curva es más achatada y se denomina curva platicúrtica, como la de la figura:

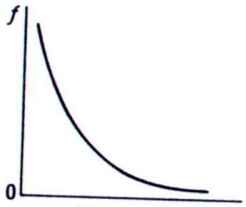


Cuando la distribución de frecuencias presenta las puntuaciones más normalmente» distribuidas, la curva no está muy apuntada ni achatada, y se llama *mesocúrtica.* En este caso, el término «normal» tiene un significado técnico muy preciso, que discutiremos más adelante. También se dice que la curva mesocúrtica, por la suavidad de sus curvas tiene forma de campana:



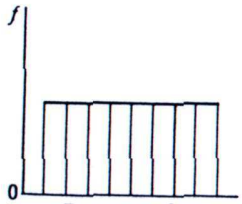
Mesocúrtica
(forma de campana)

Existen otras formas de curvas que se presentan con cierta frecuencia en el análisis estadístico de las distribuciones de frecuencias. Se denominan por aproximación a la forma global que adquieren. Así, la *curva J* responde a una distribución en la que casi todos los casos se encuentran concentrados en un extremo de la escala y, desde allí, cae uniformemente en una dirección, tal como se ve en la figura:



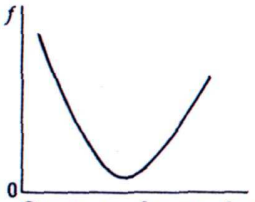
Curva J

Una distribución rectangular tiene idénticas frecuencias en todas las categorías; de ahí que su representación gráfica sea una línea paralela al eje X:



Rectangular

Finalmente, señalemos otra forma de curva que aparece con cierta frecuencia en los análisis estadísticos. Se trata de la distribución en *forma de U*, curva que aparece en las distribuciones bimodales con las modas en ambos extremos y un área de bajas frecuencias en el centro de la distribución, tal como se observa en la figura:



Curva en forma de U

3.4.2. *Medidas de la forma de una distribución. Momentos*

En la sección anterior hemos descrito la forma de una distribución haciendo referencia a conceptos generales, tales como simetría, curtosis o número de puntas, que ofrecen una imagen intuitiva y directa de dicha forma. Ahora vamos a introducir una serie de medidas o índices que, al igual que en el caso del estudio de la tendencia central, nos van a permitir fijar numéricamente las características descritas. La propia media, e incluso la mediana, o el uso de cuartiles y percentiles, pueden ser de ayuda para describir la forma de una distribución, pero existen otras medidas que son todavía de mayor utilidad.

Cuando tenemos datos medidos a nivel de intervalo resulta, con frecuencia, útil describir los datos en términos de su agrupamiento equilibrado alrededor de algún punto central. Así, por ejemplo, la media aritmética es el punto alrededor del cual el «equilibrio» algebraico de las puntuaciones es perfecto, ya que la suma algebraica de las desviaciones de las puntuaciones es cero. La desviación de las puntuaciones en relación a la media de una distribución se suele expresar mediante la letra minúscula $x = (X_i - X)$.

El momento de primer orden con respecto a la media aritmética es, simplemente, el promedio de la primera potencia de las desviaciones con respecto a la media; esto es:

$$m_1 = \frac{\Sigma x}{N} \qquad [3.18]$$

Dado que la suma de las desviaciones con respecto a la media es siempre cero, el momento de primer orden es también cero, lo que representa una característica definidora de la media. Si se utilizan potencias más elevadas, se obtienen nuevas medidas que ofrecen mayor información' estadística. Así, el *momento de segundo orden* es la varianza:

$$m_2 = \frac{\Sigma x^2}{N}$$

Otros dos *momentos* de interés estadístico son los de tercer y cuarto orden, que se definen como los promedios de las potencias de tercer y cuarto orden de las desviaciones con respecto a la media, respectivamente:

$$m_3 = \frac{\Sigma x^3}{N} \qquad [3.19]$$

$$m_4 = \frac{\Sigma x^4}{N} \qquad [3.20]$$

En general, el momento de orden *r* de una distribución de frecuencias con respecto a un origen arbitrario X_0 viene dado por la expresión:

$$m_r = \frac{1}{N}\Sigma(X_i - X_0)^r \qquad [3.21]$$

Si $X^0 = 0_r$ se tienen los momentos respecto al origen $1/N\ \Sigma x_i^r$. Con todo, los momentos más utilizados en estadística son los momentos con respecto a la media y ello por las dos ventajas que presentan. En primer lugar, por el hecho de que las potencias de orden par tienen el efecto de eliminar los signos negativos, pero las de orden impar preservan los signos negativos en el numerador de los momentos, y, en segundo lugar, por el hecho de que las potencias más altas tienden a destacar mayores desviaciones con respecto a la media.

El momento de tercer orden es un índice de asimetría porque es un momento impar: en consecuencia, si las puntuaciones altas y bajas no se equilibran alrededor de la media, no sería igual a cero. Además, como se trata de un momento elevado, acentúa las desviaciones extremas con respecto a la media que puedan existir. El momento de cuarto orden es un momento par, por lo que no diferencia entre las desviaciones por encima o por debajo de la escala media. Como se trata de un momento elevado, acentúa también las desviaciones de las puntuaciones que se encuentran en ambos extremos de la distribución. Por eso, el momento de cuarto orden resulta útil como medida del grado de curtosis en una distribución.

Los momentos vienen medidos en las unidades de medición de las puntuaciones de la distribución correspondiente. Pero como con frecuencia hacen falta medidas relativas de la asimetría y de la curtosis que no tengan en cuenta la unidad de medición, en tal caso se utilizan dos medidas, B_1 y B_2, que se definen del siguiente modo:

$$B_1 = \frac{m_3}{\sqrt{m_2^3}} \qquad [3.22]$$

$$B_2 = \frac{m_4}{m_2^2} \qquad [3.23]$$

El primero se utiliza como medida del sesgo o asimetría, y el segundo como medida de curtosis. Veamos algunas de sus propiedades.

El *sesgo* es el grado de asimetría, o falta de simetría, de una distribución. Ya hemos visto anteriormente que si la curva de frecuencias de una distribución tiene una «cola» más larga a la derecha del máximo central que a la izquierda, se dice de la distribución que está sesgada a la derecha o que tiene sesgo positivo. Si ocurre lo contrario, se dice que la curva está sesgada a la izquierda o que tiene sesgo negativo. También hemos visto con anterioridad que, según el grado y tipo de simetría, así se sitúan en orden relativo la moda, la media y la mediana. Pues bien, una forma de medir el sesgo de una curva viene dada por la siguiente fórmula:

$$\text{Sesgo} = \frac{\text{Media-Moda}}{\text{Desviación típica}}$$

Ahora bien, esta fórmula requiere el cálculo de tres índices, por lo que se utiliza una fórmula más sencilla en base a los momentos de segundo y tercer orden, y que no es otra que el coeficiente B_1:

$$\text{Coeficiente de sesgo} = B_1 = \frac{m_3}{\sqrt{m_2^3}}$$

Si la curva está sesgada a la derecha, B_1 tendrá un valor positivo, mientras que si el sesgo es negativo, B_1 ofrecerá un valor negativo. Por ser una magnitud relativa, B_1 expresa la cantidad relativa de asimetría y puede ser utilizada para comparar distribuciones que contienen diferentes unidades de medición.

En cuanto a B_2, se utiliza como coeficiente de curtosis o medida del grado de apuntamiento de una distribución. Los valores pequeños de B_2 representan una curva platicúrtica (más baja que la curva normal), mientras que valores altos de B_2 indican una distribución leptocúrtica o apuntada. La curva normal tiene un valor de B_2 igual a tres. A continuación vamos a ocuparnos de este último tipo de distribución.

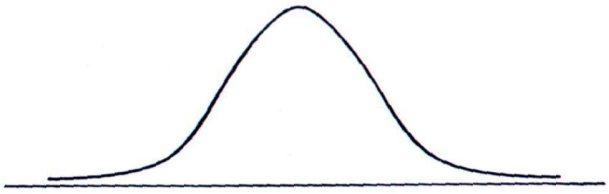
3.5. LA DISTRIBUCIÓN NORMAL.

Vamos a tratar ahora un tipo especial de distribución de frecuencias, la curva normal, que es muy importante en el análisis estadístico. Tal distribución resulta útil no sólo porque un gran número de distribuciones de frecuencias presentan formas aproximadamente normales, sino también por la significatividad teórica de la curva normal en el campo de la estadística inferencial. Ahora no vamos a ocuparnos de este último aspecto, limitándonos a exponer las propiedades de la curva normal en relación a la desviación típica. [1]

Antes de continuar adelante conviene que distingamos entre *distribuciones de frecuencias finitas y distribuciones de frecuencias infinitas*. Las distribuciones que hemos visto hasta ahora siempre se han referido a un número finito de casos. Sin embargo, resulta útil, desde un punto de vista matemático, pensar en términos de distribuciones basadas en un número infinito de casos. Tales distribuciones vendrán representadas por curvas cuyos extremos se van acercando suavemente al eje X, pero sin cruzarse con él, y que, además, pueden expresarse por medio de ecuaciones matemáticas relativamente simples. La distribución normal es una curva de este tipo. Veamos algunas de sus características.

3.5.1. *La curva normal*

La curva normal responde al tipo de curva perfectamente simétrica, basada en un número infinito de casos, por lo que sólo puede ser tratada de forma aproximada cuando se opera con datos reales. Tiene una forma acampanada, tal como se observa a continuación:



Forma general de una curva normal

Por tratarse de una curva simétrica y unimodal, coinciden la media, la moda y la mediana. La ecuación matemática de la curva normal es relativamente simple, en la que intervienen la desviación típica s y las desviaciones de las puntuaciones con respecto a la media IX-XI, de la forma siguiente:

$$Y = \frac{1}{s\sqrt{2\pi}} exp \left| -\frac{(X-\bar{X})^2}{2s^2} \right| \qquad [3.24]$$

en donde Y representa la altura de la curva para cualquier valor dado de X, valor de la puntuación en la abscisa; *exp* representa la base e de los logaritmos naturales, elevada a la potencia indicada entre paréntesis, y π es el número pi. No resulta necesario memorizar esta fórmula, sino recordar simplemente que en su composición intervienen la media y a desviación típica. Además, en la práctica nunca se utiliza la fórmula [3.24], ya que para operar con ella se utilizan unas tablas que dan directamente el área que queda por debajo de la curva normal para determinados intervalos. Esta tabla se ha podido construir basándose en una importante propiedad de la curva normal, y es que, con independencia de los valores particulares que tomen la media y la desviación típica de una curva normal cualquiera, habrá siempre un área constante (o proporción de casos) entre la media y una ordenada que se encuentre situada a una distancia dada con respecto a la media en términos de unidades de desviación típica.

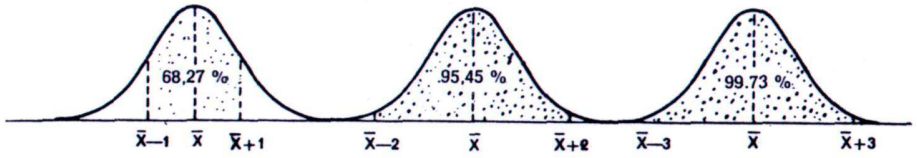
En términos estadísticos, resulta conveniente considerar una curva normal cuyas puntuaciones se expresen en puntuaciones típicas -puntuaciones *z*- en lugar de sus unidades originales (pesetas, años, etc.). Es lo que se llama una curva normal tipificada, y al venir la variable X presada en unidades de desviación $z = |X - X|/_s$, la ecuación [3.24] ceda sustituida por la forma llamada tipificada:

En este caso se dice que la curva se distribuye normalmente con media cero y varianza uno.

Un gráfico de esta curva normal tipificada se muestra en la figura 2, indicándose en el mismo gráfico las áreas incluidas entre s = -1 y + 1, s = -2 y +2, s = -3 y +3, que son, respectivamente, el 68,27, 95,45 y 99,73 por 100 del área total, que, como se recordará, vale uno.

FIGURA 2

Areas bajo la curva normal



Dicho de otra forma, alejándonos una unidad, dos unidades o tres unidades de desviación típica con respecto a la media se encuentra el 68,27, el 95,45 y el 99,73 por 100, respectivamente, del área total.

Esta propiedad de la curva normal ofrece una interpretación de la desviación típica y un método para visualizar su significado. Y es que son muchas las distribuciones de frecuencias que son lo suficientemente parecidas a la distribución normal como para que en ellas se den también las anteriores relaciones entre áreas y desviaciones típicas. Incluso en el caso de distribuciones de ingresos económicos, o distribuciones de la talla y del peso de la población, que son ligeramente asimétricas en la dirección de los valores altos, habitualmente se puede encontrar que los dos tercios de los casos se encuentran dentro de una unidad de desviación típica con respecto a la media.

Los valores numéricos de cualquier curva normal pueden transformarse de tal modo que una sola tabla puede ser utilizada para evaluar la proporción de casos que queda dentro de un determinado intervalo. Supongamos, por ejemplo, que tenemos una curva normal de media 60 y desviación típica 15, y que deseamos conocer la proporción de casos que queda dentro del intervalo 60 a 85. Para ello, calculamos en primer lugar el número de unidades de desviación típica que separa a 85 de 60, y lo hacemos mediante la fórmula:

$$z = \frac{X - \bar{X}}{s} = \frac{85 - 60}{15} = 1{,}66$$

El valor de z = 1,66 indica que la ordenada se encuentra a 1,66 unidades de desviación típica con respecto a la media. Para saber la proporción de casos que queda dentro de dicho intervalo recurriremos a la tabla B del apéndice, en la que aparecen las áreas que quedan por debajo de la curva normal para diferentes valores de z. Los valores de z, aparecen en la columna de la izquierda y en la fila superior. Los dos primeros dígitos de z se obtienen leyendo a lo largo de la columna de la izquierda, y el tercer dígito leyendo en la fila superior. Las cifras que forman el interior de la tabla indican la proporción del área entre la media (que vale 0) y la ordenada correspondiente a z. En el ejemplo anterior, con z = 1,66 el área que queda dentro de tales límites vale 0,4515. Si el valor de z hubiera sido 1,6 el área correspondiente hubiera sido 0,4452. Es decir, que aproximadamente el 45 por 100 de los casos queda dentro del intervalo 60 a 85 en la distribución normal de media 60 y desviación típica 15.

Aunque hemos dicho anteriormente que muchas distribuciones de frecuencias se asemejan a la distribución normal, son más todavía las que se alejan del modelo normal. En tal caso, no se pueden utilizar para estas distribuciones las propiedades de la desviación típica que se han visto al estudiar la curva normal. De ahí que para describir correctamente tales distribuciones habrá que recurrir a otras medidas de tendencia central, forma y variación.

3.6 TERMINOLOGÍA

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

- Posición o tendencia central de una distribución.
- Moda.
- Mediana.
- Media aritmética.
- Media geométrica.
- Media armónica.
- Media cuadrática.
- Variación o dispersión de una distribución.
- Recorrido o rango.
- Recorrido intercuartílico, recorrido semiintercuartílico.
- Desviación media.
- Desviación típica.
- Varianza.
- Coeficiente de variabilidad.
- Puntuaciones normalizadas o típicas.
- Simetría/asimetría de una distribución. Sesgo.
- Curtosis.
- Momentos de orden n.
- Distribución normal. Curva normal.

EJERCICIOS

1. Calcular la moda, mediana y media en la distribución de frecuencias del ejercicio 4 del capítulo 2.

2. Calcular la moda, mediana y media en la distribución de frecuencias del ejercicio 5 del capítulo 2.

3. En una encuesta de opinión pública la población se autoubicó en una escala ideológica izquierda-derecha (recorrido 1-10) tal como aparece en la siguiente distribución. Calcular la media y la mediana.

Escala izquierda-derecha	f_i

1– 2	13
3– 4	17
5– 6	42
7– 8	17
9- 10	11

4. Las calificaciones de un estudiante en los cuatro exámenes parciales del curso fueron 5, 7, 6, 8. Si los pesos asignados a cada examen son 1, 2, 2, 1, ¿cuál es la nota final del curso? ¿Cuál sería si todos los pesos fuesen iguales?

5. El salario medio percibido por los empleados de una empresa es 80.000 pesetas. El salario medio de un hombre en dicha empresa es 85.000 pesetas y el de las mujeres 78.000 pesetas. Determinar el porcentaje de hombres y mujeres que trabajan en la empresa.

6. Calcular el recorrido, el rango intercuartílico, la desviación media, la varianza y la desviación típica en la distribución de frecuencias del ejercicio 4 del capítulo 2.

7. Calcular el recorrido, el rango intercuartílico, la desviación media, la varianza y la desviación típica en la distribución de frecuencias del ejercicio 5 del capítulo 2.

8. Si la media de una distribución normal es 70 y su desviación típica 8:

a) ¿Qué proporción de casos se encuentra entre 70 y 85?

b) ¿Qué proporción de casos se encuentra entre 80 y 93?

c) ¿Qué proporción de casos es menor de 65?

d) ¿Cuántas unidades de desviación típica a ambos lados de la media hay que recorrer para obtener dos colas que contengan cada una de ellas el 3 por 100 del área total? ¿Y el 10 por 100?

e) ¿Qué puntuación tiene el 5 por 100 de los casos por encima de ella? (es decir, localizar el percentil 95).

9. Supóngase que una curva normal tiene una media de 50 y que el 7 por 100 de los casos tiene puntuaciones por encima de 70. ¿Cuál es la desviación típica?

BIBLIOGRAFIA

ALCAIDE INCHAUSTI, Ángel: *Estadística aplicada a las Ciencias Sociales*, Madrid, Pirámide, 1976.

AMÓN, Jesús: *Estadística descriptiva para psicólogos* Madrid, 1973,

BLALOCK, Hubert M.: *Social Statistics*, New York, McGraw-Hill, 1960.

DÍEZ NICOLÁS, H. y J. R. TORREGROSA: "Aplicación de la Escala de Cantril en España: Resultados de un estudio preliminar", *Revista Española de la Opinión Pública* 10, 1967, Págs. 77-100.

JIMÉNEZ BLANCO, José, et al.: *La conciencia regional en España*, Madrid, C.I.S., 1977.

LOETHER, H. J., y D. G. McTAVISH., *Descriptive Statistics for Sociologists*, Boston. Allyn and Bacon, 1974.

NOTAS:

* El procedimiento es el siguiente: aplicando logaritmos a la expresión [3.7] se tiene que, $\log M_G = \log \sqrt[n]{(X_1)(X_2)\ldots(X_n)} = 1/n \log [(X_1)(X_2)\ldots(X_n] = 1/n [\log(X,) + \log(X_1) + \log(X_2) + \ldots + \log(X_n)]$. Una vez calculada esta expresión, el valor de M_G se obtendrá tomando el antilogaritmo de la misma, esto es, que $M_G = \text{antilog} [1/n (\log(X_1) + \log(X_2) + \ldots + \log(X_n)]$

[1] Al estudiar las pruebas de decisión estadística y la teoría de las muestras en próximos capítulos, se hará evidente la utilidad de la distribución normal en la estadística inferencial. El objetivo de la presente sección es el de mostrar las propiedades de la curva normal y el uso de las tablas basadas en ella.