Herramientas estadísticas en criminología

Colección:
Criminología - Manuales

Coordinadores:
Cristina Rechea Alberola
Antonio Andrés Pueyo
Andrea Giménez-Salinas Framis

Herramientas estadísticas en criminología

Concepción San Luis Costas
Isabel Cañadas Osinski

EDITORIAL SINTESIS

Consulte nuestra página web: **www.sintesis.com**
En ella encontrará el catálogo completo y comentado

Vallehermoso, 34. 28015 Madrid
Teléfono: 91 593 20 98
www.sintesis.com

ISBN: 978-84-9171-396-8
Depósito Legal: M. 27.170-2019

Impreso en España - Printed in Spain

Índice

Contenidos digitales

2

Explorando los datos

2.1. Introducción

La primera cuestión que debe contemplarse, antes de la aplicación de cualquier técnica estadística, es la relativa a la descripción de la muestra (en algunos casos excepcionales de la población); la segunda, el estudio minucioso de las características que presentan las variables objeto de estudio que han sido cuantificadas mediante el correspondiente proceso de medida o asignación numérica.

El análisis exploratorio de los datos implica una actitud curiosa que está motivada por la premisa de que cuanto mejor conozca el investigador los datos que tiene, más eficientemente se pueden usar para desarrollar y refinar sus propuestas. Por tanto, el trabajo de detective del investigador tendrá como principal objetivo un abordaje exploratorio de datos para maximizar lo que pueda aprender a partir de ellos y de este modo entender el conjunto de las variables y su comportamiento.

2.2. Exploración de los datos para una variable

Realizar un análisis exploratorio de los datos permitirá al investigador, mediante un conjunto de procedimientos sencillos:

- Conocer la estructura de los datos, es decir, ver las características que presenta su distribución e intuir la distribución de probabilidad que presentan, visualizando la forma y sus propiedades características (normalidad, simetría, apuntamiento o curtosis).
- Detectar posibles errores en el diseño.
- Detectar errores en el proceso de recogida de datos.
- Identificar casos atípicos, fruto de diversas causas que pueden abarcar desde errores humanos hasta la presencia de valores anómalos en relación con el conjunto general de valores.

- Evaluar y tratar datos ausentes (datos que faltan en la matriz).
- Comprobar los supuestos subyacentes a la técnica estadística que se pretende emplear para responder a las preguntas de la investigación.

Para cumplir con estos objetivos deberá construir una distribución de frecuencias y acudir a las representaciones gráficas y numéricas, lo que se explica en las siguientes páginas.

2.2.1. Distribución de frecuencias

Una vez se tiene la matriz de datos, es el momento de empezar a organizar variables, sujetos, analizar los casos, etc. El primer paso es la construcción de una *distribución de frecuencias*: disposición de una variable en una tabla con sus modalidades, que incluye el recuento de casos en cada una, los porcentajes, etc. Hay tantas distribuciones de frecuencias como variables contenga la matriz: una distribución para cada una de las variables.

CUADRO 2.1. *Matriz de datos*

Sujeto	*Creencias*	*Motivación*	*Empatía*	*Distorsiones*	*Emociones*	*Control*	*Responsabilidad*
1	110	51	20	123	1	13	8
2	100	50	2	110	1	10	10
3	120	33	8	120	2	31	10
4	120	12	22	131	4	42	9
5	50	25	5	150	9	55	4
6	110	11	21	110	10	61	9
7	60	18	16	175	6	27	8
8	30	30	3	133	3	33	5
9	40	54	24	120	4	40	4
10	90	41	29	149	7	90	7

La matriz de datos del cuadro 2.1 es un ejemplo de los resultados obtenidos, tras un programa de intervención a hombres maltratadores, en la aplicación de una batería de cuestionarios para medir sus creencias sexistas, su motivación para el cambio, la empatía con las víctimas, las distorsiones cognitivas, la identificación de emociones, el control de la ira y la asunción de responsabilidad. En la matriz de datos están los sujetos con sus puntuaciones en las variables, pero la visión de estas resulta un tanto confusa, al no estar ordenadas sus puntuaciones.

Al extraer, por ejemplo, las variables creencias sexistas y asunción de responsabilidad y organizar cada una en una distribución de frecuencias (cuadro 2.2), se puede

observar todo el rango de valores y el número de casos (n_i) de sus modalidades. Enseguida se aprecia, por ejemplo, que las puntuaciones en creencias se concentran más en torno a los valores más altos de la distribución (hay 6 sujetos con valores entre 90 y 120), casi lo mismo que la responsabilidad, cuyo mayor número de casos se encuentra en los valores más altos (6 sujetos entre 8 y 10), aunque, en general, sus puntuaciones estén más repartidas.

CUADRO 2.2. *Distribuciones de frecuencias*

Creencias	n_i	*Responsabilidad*	n_i
30	1	4	2
40	1	5	1
50	1	6	0
60	1	7	1
70	0	8	2
80	0	9	2
90	1	10	2
100	1		10
110	2		
120	2		
	10		

La ventaja fundamental de una distribución de frecuencias es la posibilidad de extraer conclusiones de la variable sin realizar siquiera un análisis estadístico complejo. Incluso, sin perder sencillez, se puede complementar con otros datos:

- *Frecuencia absoluta* (n_i): número de casos de cada modalidad de la variable.
- *Frecuencia acumulada* (n_a): recuento de las frecuencias absolutas en orden ascendente según el sentido de las modalidades.
- *Proporción o frecuencia relativa* (p_i): frecuencia absoluta / número total de casos (n).
- *Proporción acumulada* (p_a): frecuencia acumulada (n_a) / número total de casos (n).
- *Porcentaje* (P_i): proporción (p_i) x 100.
- *Porcentaje acumulado* (P_a): proporción acumulada (p_a) x 100.

En el caso de la variable creencias sexistas quedaría tal y como se aprecia en el cuadro 2.3. En general, sus valores se encuentran bastante repartidos en la distribución, aunque el mayor porcentaje de casos se sitúa en los valores 110 y 120 y ningún sujeto ha obtenido una puntuación de 70 u 80.

CUADRO 2.3. *Distribución de frecuencias de la variable creencias*

Creencias	n_i	n_a	p_i	p_a	P_i	P_a
30	1	1	0,1	0,1	10	10
40	1	2	0,1	0,2	10	20
50	1	3	0,1	0,3	10	30
60	1	4	0,1	0,4	10	40
70	0	4	0	0,4	0	40
80	0	4	0	0,4	0	40
90	1	5	0,1	0,5	10	50
100	1	6	0,1	0,6	10	60
110	2	8	0,2	0,8	20	80
120	2	10	0,2	1,0	20	100
	n = 10	–	1,0	–	100	–

Cabe decir que la frecuencia, proporción y porcentaje acumulados (n_a, p_a y P_a) solo se utilizan con variables cuantitativas. Pensando, por ejemplo, en la variable estado civil, con sus modalidades y frecuencias absolutas, ¿qué sentido tendría hacer un recuento acumulado de sus frecuencias? Aunque ahora no se aprecie la utilidad de la acumulación, más adelante se comprobará lo práctico que resulta para localizar sujetos de la distribución y para interpretar su posición.

2.2.2. Representaciones gráficas. Valores perdidos y atípicos

Una manera sencilla de observar el comportamiento de una variable es mediante su representación gráfica, no en vano, su función principal es obtener informaciones globales mediante un solo golpe de vista. Existen gráficas para todos los gustos, pero no siempre su aspecto sofisticado es sinónimo de claridad y exactitud. Las representaciones más sencillas son las que mejor reflejan las características de las variables y estas son las que se van a presentar a continuación.

Cuando la variable es cualitativa o cuantitativa discreta, un gráfico muy utilizado es el *diagrama de barras*. En la abscisa se colocan las modalidades de la variable y en la ordenada las frecuencias o porcentajes. También se puede representar la variable medida en dos o más grupos para su comparación, tal y como se puede ver en la figura 2.1.

En el diagrama de la figura 2.1 se compara, por ejemplo, el grado de ira en un grupo de hombres y otro de mujeres. Las diferencias en las modalidades *baja, media, alta* y *muy alta* no están muy claras. Se podría decir que en las tres se presentan niveles similares de ira.

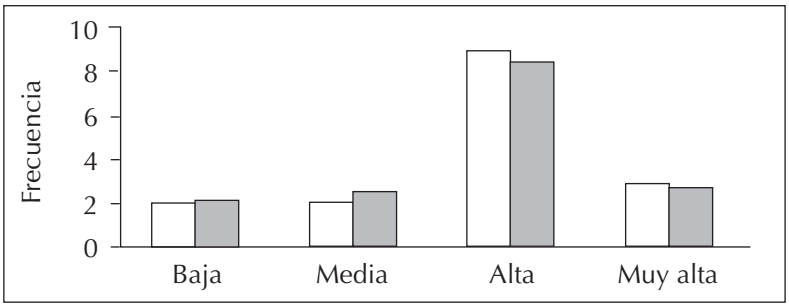

FIGURA 2.1. Diagrama de barras.

Cuando se trata de variables cuantitativas, los diagramas se denominan *histogramas*, como se muestra en la figura 2.2. En este ejemplo se observa la evolución de un grupo de adolescentes diagnosticados de un trastorno adaptativo, antes y después de una intervención para mejorar su grado de incertidumbre en temas relacionados con la identidad (objetivos a largo plazo, elección profesional, patrones de amistad, orientación y conducta sexual, etc.). En el primer gráfico todos los sujetos tienen una puntuación superior a 35, situándose un gran número de ellos entre los valores 40 y 50 de la escala. En el segundo gráfico es donde se constata la mejoría tras el tratamiento, puesto que ningún sujeto supera el valor de 50 en la escala de incertidumbre, ubicándose un elevado número de ellos entre los valores 10 y 30. Se trata de dos histogramas bien sencillos que arrojan mucha información. Es importante señalar que se debe respetar el origen de coordenadas en ambos gráficos, si no, resultan difíciles de comparar y pierden toda su función de informar rápidamente de los cambios en una variable, de su comportamiento en grupos distintos, etc.

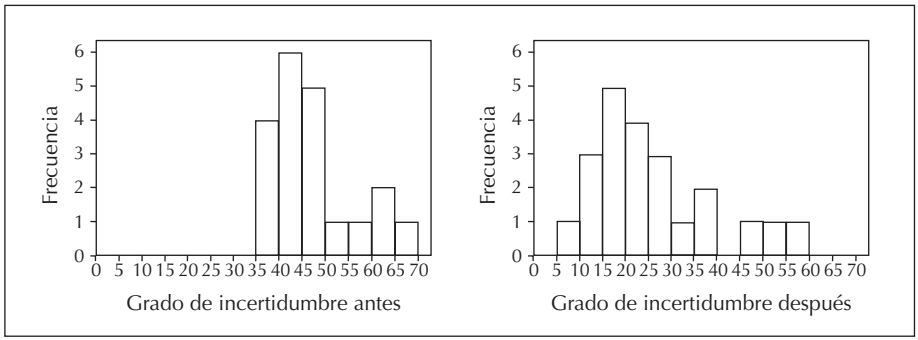

FIGURA 2.2. Histograma de la misma variable en dos momentos distintos.

Un gráfico digno de mención es la *curva normal*, que es el modelo prototipo contra el que se suelen comparar la mayoría de las variables cuando se trabaja con métrica cuantitativa. Se caracteriza por ser simétrica y presentar una línea lisa o suavizada con dos colas de igual longitud, lo que implica que el porcentaje de casos en ambas es idéntico. En realidad, se trata de un histograma cuando se han recogido infinitos datos de la variable; en otras palabras, se trata de un modelo teórico.

Sin embargo, y pese a ser este el modelo por antonomasia, no todas las variables se acercan a esta distribución; es más, en muchas ocasiones es extremadamente complicado, siquiera, encontrar variables que se asemejen a él. La situación más habitual es que cada una de las variables que se haya medido presente diferencias tanto en localización como en forma y dispersión respecto al modelo representado en la figura 2.3, de ahí la importancia de describir para cada variable estas características, lo que permitirá evaluar el grado de semejanza con el modelo teórico sobre el que descansará el análisis de datos.

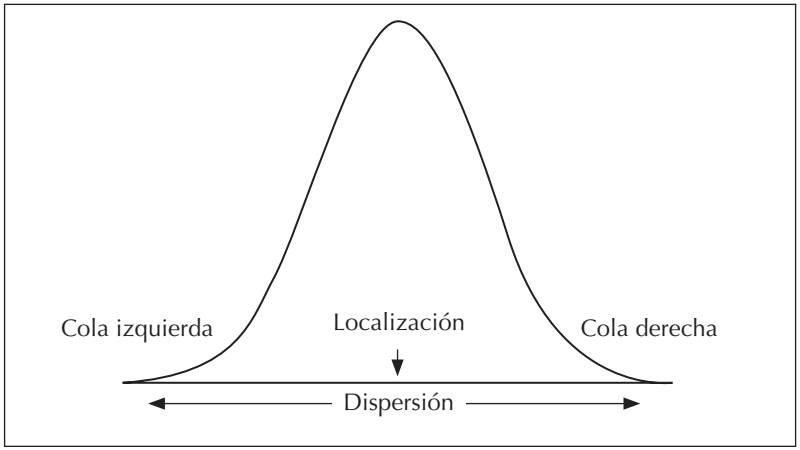

FIGURA 2.3. Curva normal.

Un aspecto que condicionará en gran medida la obtención de representaciones numéricas son los casos raros o atípicos. Un valor raro hace referencia a un sujeto que presenta un comportamiento muy diferente del resto, normalmente, porque tiene puntuaciones extremas por encima o por debajo de la cola de la distribución, lo que provoca que esta se aleje de una normal. Una de las consecuencias de la presencia de *outliers*, por ejemplo, es el desplazamiento artificial de la media hacia arriba o hacia debajo, convirtiéndolos en observaciones influyentes (Kaplan, 2000).

La presencia de estos datos puede deberse simplemente a errores de codificación. Bajo tal circunstancia son fácilmente detectables mediante un análisis de frecuencias

que permita identificar valores fuera del rango admisible. En cualquier caso, los valores raros son fáciles de identificar: mediante una simple inspección visual de la representación gráfica se observan los sujetos cuyas puntuaciones están muy por encima o por debajo de la cola de la distribución o del intervalo establecido. Una vez identificado el sujeto en la base de datos, habría que eliminarlo y volver a repetir los análisis. Cuando son detectados, hay que de plantearse si esos sujetos proceden de otra población distinta o si, por el contrario, perteneciendo a la población de referencia se trata de sujetos excepcionales. El problema es que no siempre es fácil hacer esta distinción.

Por su parte, los *valores perdidos* se deben a una gran variedad de circunstancias que pueden ocurrir durante una investigación y que están fuera del control del investigador. Así, por ejemplo, puede suceder que haya fallos en los equipos (registros psicofisiológicos, por ejemplo) o que los sujetos no respondan deliberadamente a determinadas cuestiones, o bien, por muerte experimental.

Ante los valores perdidos se deben considerar su cantidad y el grado en que afectan a la investigación. En este caso, no hay una respuesta clara y sí una gran variedad de opiniones. Cohen y Cohen (1983) entienden que es soportable hasta un 5 o 10 por ciento de casos perdidos. En segundo lugar, se ha de valorar si la estructura de los datos perdidos es aleatoria o sistemática. Si la distribución de los valores perdidos es aleatoria, estaría indicando que los sujetos en los que hay presentes valores perdidos difieren solo por azar de aquellos en los que no están presentes. Por tanto, los resultados obtenidos en estos sujetos podrían ser generalizables al resto. En el caso de que la distribución de datos perdidos siga un patrón sistemático se trataría del caso contrario y, por tanto, no se debería ignorar esta circunstancia. Es más, se debería hacer todo lo posible por intentar interpretar el comportamiento y naturaleza de la distribución de los valores perdidos (por qué y cómo se distribuyen a lo largo de las distintas variables y de los sujetos).

2.2.3. Representaciones numéricas: índices de tendencia central, de variabilidad y de posición

En muchos casos, una distribución de frecuencias y una representación gráfica serán más que suficientes para entender el comportamiento de una variable en un grupo de sujetos. Sin embargo, cuando el interés se centra en la comparación de varios grupos en la misma variable, o en más de una, tratar con tablas y gráficos y más tablas y más gráficos puede resultar un poco engorroso, tanto para el propio manejo como para el lector de los informes. En situaciones así, puede ser más apropiado trabajar con unos índices que describan de forma más sencilla la información contenida en las distribuciones. Se trata de los *estadísticos* de una muestra, índices que, además de hacer manejable la información contenida en los datos, son la base sobre la que se sustenta la mayoría de las técnicas de análisis de datos.

Los *estadísticos de tendencia central* son aquellos que resumen el comportamiento de una variable en un solo valor, representando así al conjunto de sujetos en el que son calculados.

La *media aritmética* es el índice de tendencia central más utilizado y su cálculo es tan simple como promediar todas las puntuaciones de un grupo en una variable, tal y como reflejan sus fórmulas para datos brutos y para datos de una distribución de frecuencias:

$$\bar{X} = \frac{\sum X_i}{n} \qquad \bar{X} = \frac{\sum n_i \cdot X_i}{n}$$

donde X_i es la puntuación de un sujeto; n es el tamaño de la muestra y n_i la frecuencia absoluta.

Sea el siguiente ejemplo. En el servicio de psiquiatría de un centro penitenciario se ha observado que los internos drogodependientes con insomnio crónico muestran cierta sensibilidad inusual al ruido. Debido a las dificultades que conlleva la aplicación de cuestionarios que midan la sonoridad subjetiva, se decide evaluarlos con una prueba en la que deben interrumpir un sonido cuando les resulta molesto y se recoge el tiempo (en segundos) que tardan en pulsar el interruptor. Los datos de diez pacientes son los siguientes:

4 5 8 2 4 2 5 2 2 6

La media aritmética es:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{4+5+8+2+4+2+5+2+2+6}{10} = \frac{40}{10} = 4$$

Suponiendo que en los sujetos sin la dolencia descrita la media aritmética es 8 segundos, se puede concluir que estos pacientes, que han tardado en conjunto un promedio de 4 segundos en sentir el sonido como un ruido molesto, presentan una sensibilidad al ruido mayor.

Al agrupar los datos anteriores en una tabla (véase el cuadro 2.4), la media aritmética que resulta es:

$$\bar{X} = \frac{\sum n_i \cdot X_i}{n} = \frac{8+0+8+10+6+0+8}{10} = \frac{40}{10} = 4$$

CUADRO 2.4. *Cálculo de la media en una distribución de frecuencias*

X_i	n_i	$n_i \cdot X_i$
2	4	8
3	0	0
4	2	8
5	2	10
6	1	6
7	0	0
8	1	8
–	10	40

Tal es la sencillez de la obtención de la media que se ha convertido en un estadístico omnipresente en todo informe de investigación que trate con variables. Sin embargo, antes de su cálculo deberán tenerse en cuenta las siguientes consideraciones:

- Es muy sensible a cualquier variación en los datos. Por ejemplo, un error en la introducción de un valor numérico puede resultar desastroso al calcularla:

  3 5 7 4 → $\overline{X} = 4{,}75$
  3 5 7 40 → $\overline{X} = 13{,}75$

- No se debe utilizar ante conjuntos de datos en los que hay casos extremos porque deja de cumplir con su función, que es la de representar al conjunto:

  X: 1, 2, 2, 4, 4, 8, 9, 10 cuya media es $\overline{X} = 5$

  ¿A quién representa esta media si ni siquiera su valor está en la distribución? Por esto, siempre deberá ir acompañada de un gráfico o, bien se debe informar que los datos se ajustan a una distribución simétrica.
- Con variables cualitativas no tiene sentido calcularla: ¿se puede afirmar que la media aritmética del estado civil es 2,5? También hay que ser cautelosos con el resultado de algunas variables cuantitativas discretas como, por ejemplo, el número de delitos. Se puede decir que un delincuente ha cometido el doble de delitos que otro, pero ¿es pertinente hablar de un promedio de 3,5 delitos?

La simplicidad en la obtención de la media aritmética la ha convertido en la estrella de los estadísticos de tendencia central. Doctos y legos la conocen y utilizan, pero hay

que ser cautos porque los datos no están exentos de trampas. En aquellos casos en los que no sea adecuado calcularla, se puede probar con la moda o con la mediana.

La *moda* es un índice de tendencia central que representa el valor cuya frecuencia absoluta es la más alta en la distribución. Puede utilizarse con cualquier tipo de variable, ya sea cualitativa o cuantitativa, y su obtención es muy sencilla. Los datos simplemente se ordenan y se ve cuál se repite más. Ese valor es, por tanto, la moda de la distribución. En el ejemplo de los pacientes con delirium por abstinencia, la moda, tanto con los datos brutos como agrupados en la distribución de frecuencias (cuadro 2.4) es 2 (y no 4, cuidado con esto). Este resultado se interpreta como que la mayoría de los pacientes de este grupo tarda 2 segundos en sentir el ruido como molesto. (Cabe señalar que la moda no coincide con la media, lo que ya habla de una distribución asimétrica).

Puede suceder que una distribución tenga dos valores con la máxima frecuencia, la misma o muy similar, pero que ambas se diferencien bastante del resto. En estos casos, se dice que la distribución presenta dos modas o que es *bimodal*:

2, 2, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, 8, 8, 10, 10, 10, 10, 10, 10, 10

En los datos anteriores, se tienen dos modas: $Mo_1 = 6$ y $Mo_2 = 10$

Cuando los valores de la moda son adyacentes, algunos analistas toman como estadístico la media de ambas. Sin embargo, esto no debe hacerse, puesto que el resultado puede carecer de sentido: podría tratarse de un valor que no existe en la distribución. Por otro lado, ¿cuál es el problema de presentar dos modas? Precisamente lo anterior, la obtención (o presentación) de dos modas permite ver hasta qué punto el cálculo de una media es lícito. En efecto, si las dos modas son adyacentes y centradas, no sería ningún problema, pero si no lo son, esto debe advertir de que no hay que calcular la media.

La *mediana* es un índice que informa del dato de la distribución que la divide en dos partes iguales y se define, por tanto, como aquel valor que deja por encima y por debajo el 50 por ciento de las frecuencias. Puede utilizarse con variables cualitativas cuyas modalidades puedan ordenarse y con variables cuantitativas.

Su obtención es enormemente sencilla: únicamente hay que localizar, en la columna de proporciones acumuladas de la tabla de la distribución, el valor que comprenda el 0,5 y ver a qué puntuación de la variable corresponde. En el ejemplo (cuadro 2.5), tras obtener las proporciones acumuladas, la mediana se localiza en la proporción acumulada 0,6, porque es la que incluye 0,5 (la anterior se queda en 0,4), por lo que la mediana es $Md = 4$; es decir, el valor de la distribución que divide a los sujetos en dos grupos.

CUADRO 2.5. *Localización de la mediana*

X_i	n_i	p_i	p_a
2	4	0,4	0,4
3	0	0,0	0,4
4	2	0,2	0,6
5	2	0,2	0,8
6	1	0,1	0,9
7	0	0,0	0,9
8	1	0,1	1,0
–	10	1,0	

Cabe señalar que si su valor coincide con el de la media y el de la moda, o son muy similares, la distribución es simétrica o muy cercana a la simetría. En cambio, si los resultados para los tres índices son muy diferentes, hay que tener cuidado con la forma de la distribución y lo que se ha calculado. Hay que recordar que una representación gráfica es un complemento perfecto en los análisis para aclarar los resultados.

Los índices de posición son todos aquellos que permiten situar a un sujeto dentro de una distribución de frecuencias. Su nombre genérico es *cuantiles* y los más utilizados en este contexto son los percentiles y los cuartiles.

Los *percentiles*, también llamados *centiles*, son las puntuaciones de la variable que dividen la distribución en cien partes iguales. Hay, por tanto, 99 percentiles. La nomenclatura utilizada es P_k y su valor se interpreta como la puntuación que deja por debajo de sí al k por ciento de los sujetos. Por ejemplo, si $P_{20} = 9$, se dice que 9 es el valor de la distribución que deja por debajo de sí al 20 por ciento de los sujetos.

Por su parte, los *cuartiles* son las puntuaciones de la variable que dividen la distribución en cuatro partes iguales. Hay, por tanto, tres cuartiles. La nomenclatura utilizada es Q_k (k = 1, 2 o 3) y su valor se interpreta como la puntuación que deja por debajo de sí al 25 k por ciento de los sujetos. Por ejemplo, si $Q_3 = 8$, se dice que 8 es el valor de la distribución que deja por debajo de sí al 75 por ciento de los sujetos.

Obtener percentiles o cuartiles es tan sencillo como hacerlo con la mediana: se busca en la columna de las proporciones acumuladas, o de los porcentajes acumulados de la distribución, la proporción o porcentaje de interés y se localiza la puntuación de la variable a la que pertenece.

El cuadro 2.6 es el resultado de aplicar un test que mide conductas sádicas. Un sujeto que, por ejemplo, obtiene una puntuación igual a 15 resulta que se sitúa en el percentil P_{77} y en el cuartil Q_3, lo que lo coloca en los valores altos de la distribución. Es claramente necesaria una rápida intervención.

Los percentiles y los cuartiles permiten interpretar la situación de un sujeto en la distribución de la variable que se ha medido, sin necesidad de conocer ni la puntuación directa ni el valor mínimo y máximo de la distribución, lo que resulta de gran utilidad.

Por otro lado, del cuadro 2.6 se desprende que la mediana se localiza en el percentil 50 y en el cuartil 2, siendo su valor igual a 12.

En efecto, en distribuciones simétricas:

$$Md = P_{50} = D_5 = Q_2$$

CUADRO 2.6. *Localización de percentiles y cuartiles*

X_i	n_i	n_a	p_a	P_k	Q_k
6	8	8	0,08	P_7	–
7	4	12	0,12	P_{11}	–
8	8	20	0,20	P_{19}	–
9	5	25	0,25	P_{24}	–
10	7	32	0,32	P_{31}	Q_1
11	10	42	0,42	P_{41}	–
12	*9*	*51*	*0,51*	P_{50}	Q_2
13	11	62	0,62	P_{61}	–
14	9	71	0,71	P_{70}	–
15	7	78	0,78	P_{77}	Q_3
16	8	86	0,86	P_{85}	–
17	4	90	0,90	P_{89}	–
18	4	94	0,94	P_{93}	–
19	6	100	100	P_{99}	–

No todos somos iguales. Bien es cierto que entre las personas existen ciertos comportamientos similares, pero siempre hay diferencias individuales. Mientras que los procesos como grupo vienen cuantificados con los índices de tendencia central (media, mediana, moda, etc.), las diferencias individuales se verán reflejadas en los *índices de variabilidad.* Dicho con otras palabras, estos índices permitirán saber cuánto se distancian los individuos del sujeto promedio. En el ejemplo que se presenta se recogen las puntuaciones de cinco criminales obtenidas en egocentrismo ilimitado:

X_1: 10 11 12 13 14

La media calculada es $\overline{X} = 12$. Sin embargo, los registros muestran diferencias entre ellos, por lo que cada uno se distancia de su media:

$10 - 12 = -2$; $11 - 12 = -1$; $12 - 12 = 0$; $13 - 12 = 1$, y $14 - 12 = 2$

Los resultados muestran alejamientos positivos y negativos (unos son más egocéntricos y otros menos, con respecto a la media de 12), con lo que se podría plantear el cálculo del promedio de esas distancias como índice de variabilidad. En efecto, esto es precisamente lo que hace el estadístico *desviación media*:

$$DM_{X_1} = \frac{\sum (X_i - \bar{X})}{n}$$

Su fórmula claramente refleja cuánto se alejan en promedio las puntuaciones de su media. Aplicándola a los datos del ejemplo:

$$DM_{X_1} = \frac{\sum (X_i - \bar{X})}{n} = \frac{(-2) + (-1) + 0 + 1 + 2}{5} = 0$$

Habiendo constatado las diferencias individuales, ¿cómo es posible que el promedio de distancias sea igual a cero? La razón es muy sencilla: las diferencias son unas positivas y otras negativas, compensándose las unas con las otras, y siempre va a ocurrir que la suma de esas diferencias sea igual a 0, ya que son distancias a un punto central, que es la media.

Entonces, si la desviación media siempre va a ser igual a cero, hay que buscar otro índice que, siendo tan sencillo e informativo como este, solvente el problema. Una posibilidad sería trabajar con las diferencias en valores absolutos, pero estos resultan poco manejables matemáticamente y, por tanto, poco convincentes.

Otra opción es calcular las diferencias, elevarlas al cuadrado y a partir de ahí, obtener su promedio. Un índice de dispersión computado así se denomina *varianza*:

$$S_X^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

Aplicándolo a los datos:

$$S_{X_1}^2 = \frac{\sum (X_i - \bar{X}_1)^2}{n} = \frac{4 + 1 + 0 + 1 + 4}{5} = 2$$

Una fórmula más rápida de obtener la varianza, sobre todo cuando se manejan muchos datos, es mediante la siguiente expresión: