Interpretar un caso (univariable)

Vicente Manzano Arrondo – 2012,2013

Tengo interés por saber qué piensan los jóvenes andaluces sobre el papel que pueden desempeñar en los asuntos del bien común o de la vida pública o política. Para ese estudio, hemos pensado que una variable pertinente es el número de actividades de acción política no convencional en las que se ha participado (asistir a manifestaciones, participar en recogidas de firmas, organizar con un colectivo una acción social, redactar un manifiesto, etc.). Se trata de una variable cuantitativa. Pongamos, por ejemplo, que hemos entrevistado a 50 jóvenes y hemos recogido los siguientes datos respecto a esa variable:

Datos									
1	2	3	6	5	0	4	3	1	9
7	6	6	10	7	13	20	2	3	6
5	12	2	2	1	15	11	11	0	2
1	2	4	4	5	7	5	6	10	3
2	1	8	12	8	1	0	9	14	8

El primer acercamiento es conocer la variable en términos grupales. Para ello realizamos la siguiente tabla de frecuencias (dispuesta en sentido horizontal) y su correspondiente representación gráfica.

Tabla de frecuencias																	
Xi	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
fi	3	6	7	4	3	4	5	3	3	2	2	2	2	1	1	1	1

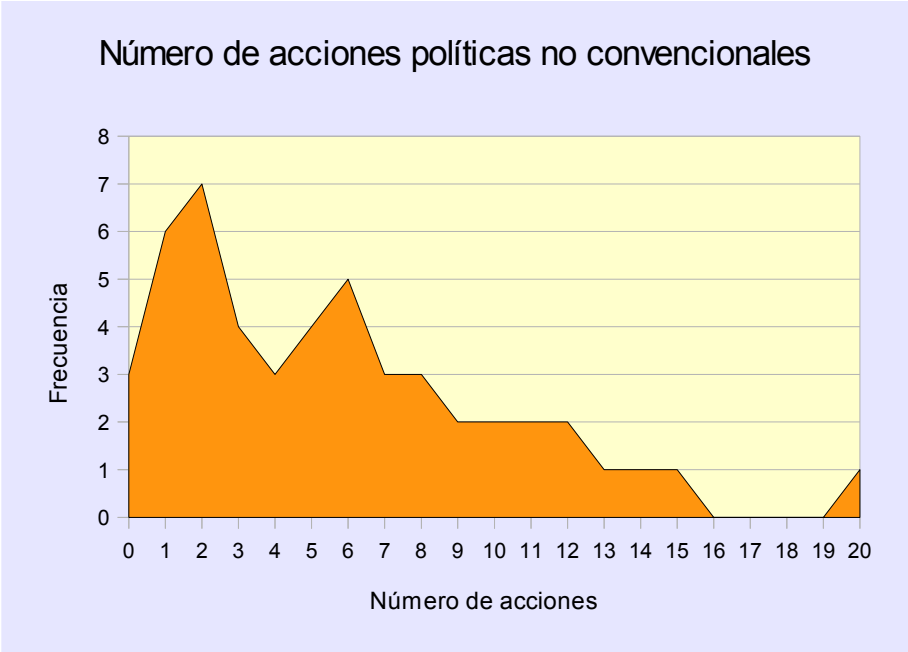

Figura 1. Diagrama de áreas.

Se puede observar que existe cierto agolpamiento de datos en los valores inferiores, y cierta dispersión hacia los superiores. En otras palabras, la mayoría de estas cincuenta personas han participado en contadas ocasiones en acciones políticas de tipo no convencional. Esta información puede complementarse con el cálculo de los

estadísticos de tendencia central o de representación numérica a los que acudimos en el caso de variables cuantitativas:

Estadísticos	
n	50
Media	5,7
Des.tip.	4,5
CV	78

La media es 5,7. Teniendo en cuenta que el centro del recorrido (de 0 a 20) es 10, vemos que el valor de la media se encuentra sensiblemente desplazada hacia los valores inferiores, lo que resalta el agolpamiento ya mencionado. La desviación tipo es muy elevada (un 78% de la media, según muestra el coeficiente de variación de Pearson). No debe asombrar. Esto ocurre siempre en distribuciones con una asimetría evidente, como pasa en este caso.

El conocimiento de la distribución de datos es insuficiente. Es habitual que nos preocupe otro aspecto: cómo se comportan los casos concretos, cómo interpretar puntuaciones concretas a la luz del conjunto en el que están insertas.

En psicología, por ejemplo, es muy habitual que nos interese la excepción al menos tanto como nos interesa la norma. Nos preocupa el modo en que alguien se aleja de lo considerado previamente más que el modo en que cumple con las expectativas. Observemos, por ejemplo, que el conjunto de datos muestra un caso que sobresale por la parte de los valores altos: una persona que dice haber participado en 20 acciones políticas no convencionales. No hay nadie con 19, ni 18, ni 17, ni 16. La siguiente persona dice haber realizado 15 de este tipo de acciones. Luego, el caso de $X_i = 20$ es muy llamativo y merece atención en dos sentidos. Nos interesa comprender no solo qué explicaciones justifican ese agolpamiento en torno a los valores bajos sino también cuáles permiten entender esas excepciones a la norma. Con estos datos no tenemos información para responder a estas inquietudes, pero tal vez el estudio que estamos realizando (donde se han medido muchos otros aspectos) nos permita arrojar luz sobre estos interrogantes.

Otra consecuencia de la existencia de estos valores peculiares es que tienen mucho efecto en las conclusiones sobre el grupo. Si volvemos a calcular los mismos estadísticos, pero eliminando el caso $X_i = 20$, la media disminuye 3 décimas y la desviación tipo disminuye 5 décimas. Puede no parecer mucho, pero recuerda que hablamos del efecto de un solo dato entre 50 y que la escala (en torno al valor 6 de media) maneja pocas unidades. Si prescindimos también de los tres casos extremos con valores $X_i = 13$, 14 y 15, entonces la media disminuye 6 décimas y la desviación tipo 5. Observa el efecto:

Estadísticos	
n	50
Media	5,7
Des.tip.	4,5
CV	78

Estadísticos	
n	49
Media	5,4
Des.tip.	4,0
CV	74

Estadísticos	
n	46
Media	4,8
Des.tip.	3,5
CV	71

Los casos peculiares, raros o extremos tienen mucho efecto en las medidas de representación numérica. La moraleja es clara: a la hora de construir una norma no podemos tener en cuenta los casos extremos.

El estudio de los casos particulares no está justificado solo porque tienen importantes efectos en las conclusiones sobre el grupo. También merecen una atención per se. En psicología, por ejemplo, habitualmente los casos son personas. Cada persona

es un fin en sí mismo, no un instrumento. En la práctica de la psicología nos preocupa el bienestar de la gente, tanto en grupo como individualmente. Si algunas personas muestran comportamientos peculiares, que no se adaptan a la norma y que no pueden recibir una atención estandarizada, es importante saber qué ocurre y obrar en consecuencia. Una misma terapia no funciona igual con todo el mundo. Un mismo método de asesoramiento educativo tampoco. Una misma estrategia de aprendizaje puede dar resultados muy diferentes en diferentes participantes.

En definitiva, hay dos razones por las que nos interesa mucho tener recursos para estudiar los casos raros o peculiares:

1. Su presencia dificulta construir la norma, encontrar el comportamiento general.
2. Son casos reales que requieren una atención específica.

Vamos a considerar dos herramientas para identificar e interpretar datos con valores que se alejan del centro o de lo característico de la distribución de datos. Vemos dos estrategias para concretarlos. La primera, que llamamos *medidas de posición*, dividen al conjunto ordenado de datos en un número concreto de partes con igual cantidad de datos, facilitando la interpretación del caso al identificar en qué parte se encuentra. La segunda, que llamamos *distancias estandarizadas*, realizan una transformación cuantitativa del valor original para que esa transformación exprese en qué grado ese caso se distancia de un valor representativo del conjunto.

Medidas de posición

Conocemos la mediana. Es el valor del punto central de la distribución ordenada de datos. Es decir, si colocamos un dato tras otro, ordenados según su valor, y buscamos la posición o punto del centro, la mediana es el valor de ese punto. Ya lo hemos visto y calculado en el documento “Conocer una variable”. La mediana permite interpretar un caso, de forma poco precisa: podemos decir si ese caso se encuentra por encima o por debajo de la mediana, es decir, si está respectivamente en la mitad superior o en la mitad inferior de los datos. En el ejemplo anterior, la mediana tiene el valor

$$Md = d_{n+1/2} = d_{25,5} = \frac{d_{25} + d_{26}}{2} = \frac{5 + 5}{2} = 5$$

Podemos interpretar un caso sabiendo si su valor está por encima o por debajo de 5. La figura 2 muestra los datos del ejemplo en un diagrama de barras donde cada dato representa un bloque. He coloreado cada bloque en función de si se encuentra por debajo o por encima de la mediana. A su vez, cada bloque está numerado para expresar si su posición es la 1 (por debajo de la mediana) o la 2 (por encima).

La figura 2 es útil también para observar los problemas irresolubles de imprecisión cuando se divide al conjunto de datos en partes. Como tenemos 50 datos, cada mitad está compuesta por 25. Al considerar 25 datos por debajo y 25 por encima de un punto, este punto se encuentra en la columna de los datos situados sobre el valor 5. Pero 2 bloques (datos) pertenecen a una mitad y otros dos bloques a la otra. Hay cuatro datos con el valor 5. Es arbitrario considerar a dos de ellos como pertenecientes a la parte 1 y los otros dos como pertenecientes a la parte 2, pues los cuatro coinciden en el mismo valor y, por tanto, han de ser interpretados del mismo modo. Pero no hay forma de solucionar esto. Fíjate que, de los 50 datos, 23 están por encima de 5, no 25. 23 están por debajo, y 4 coinciden. Esto ocurrirá con frecuencia. Por eso decimos que la mediana es el valor de un punto (la posición 25,5) y que es ese punto y no el valor necesariamente, lo

que divide al conjunto de datos en dos partes con igual frecuencia. Destaco esto aquí porque esta discrepancia entre valores y puntos aumenta cuando decidimos aumentar el número de partes en las que separar o dividir un conjunto de datos.

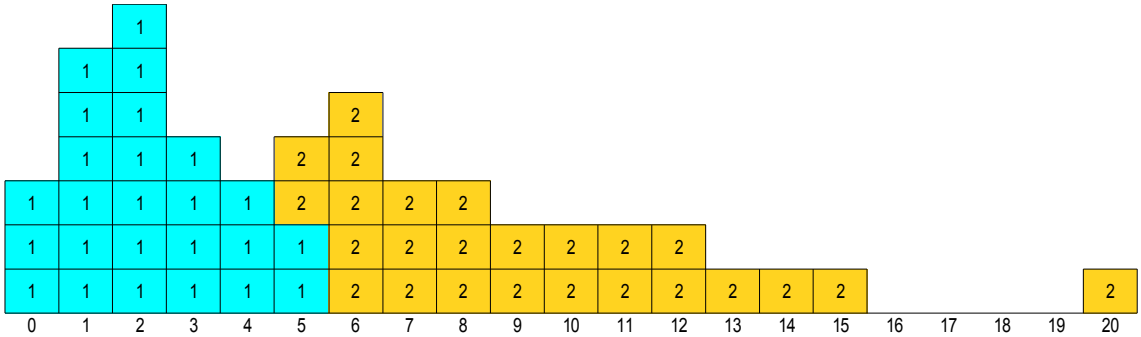

Figura 2. Coloreado y numerado de los dos lados de la mediana.

Utilizar la mediana como recurso para distinguir la posición de un caso es tan comprensible como poco útil. Le falta precisión. Sería mejor considerar más partes o posiciones para interpretar de forma más fina o precisa un dato. Las dos mitades del conjunto de datos se han conseguido gracias a utilizar un punto de corte. Pues bien, un recurso habitual es dividir la distribución ordenada de datos en cuatro partes, gracias a utilizar tres puntos de corte llamados *cuartiles* y representados mediante la letra Q con subíndice. La figura 3 muestra el resultado a partir del conjunto de datos del ejemplo.

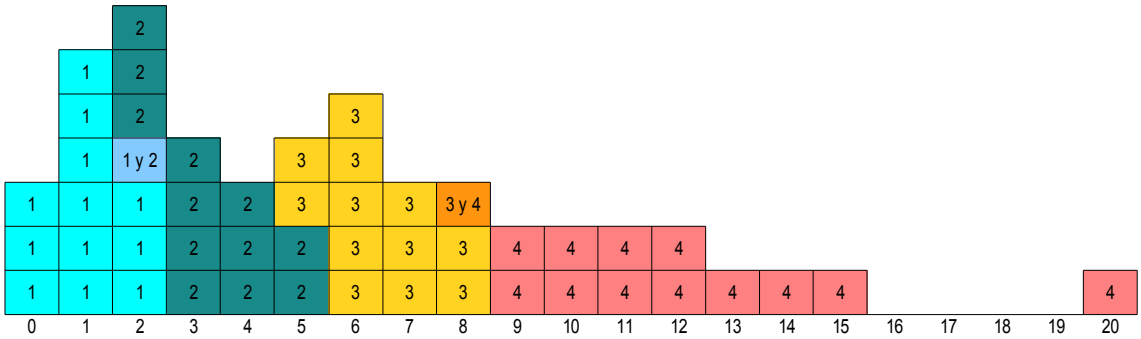

Figura 3. Coloreado y numerado de cuartiles.

Como la mitad y las dos cuartas partes son lo mismo, el segundo cuartil coincide con la mediana (Q_2 = Md). El primer cuartil, es decir, el punto de corte que distingue entre el primer y el segundo de los cuartos, se encuentra en el valor 2 (Q_1 = 2), mientras que Q_3 = 8. En el ejemplo aparece un nuevo problema. Contamos con 50 datos. La mitad son 25. Pero la cuarta parte ya no coincide con un número entero: 50/4 = 12,5. Las frecuencias con decimales no tienen sentido. Existe el dato de la posición 12 y el de la posición 13, pero no hay un dato en la posición 12,5 porque en sentido estricto esa posición no existe. Así que el dato que se encuentra en la posición 12 pertenece al primer cuarto, es decir, se encuentra por debajo del primer cuartil. El dato de la posición 14 está claramente por encima del primer cuartil y por debajo del segundo, es decir, pertenece al segundo cuarto. Pero ¿y el dato de la posición 13? La mitad de ese dato se encuentra en el primer cuarto, mientras que la otra mitad pertenece al segundo. Esto no es posible en la práctica. Es más, como ocurrió en el caso de la mediana, es arbitrario afirmar que ese dato concreto es uno u otro entre los siete datos con valor 2. Observa que tres de esos datos se encuentran en el primer cuarto, otros tres en el segundo y uno de los siete está a medio camino entre ambos cuartos. Y es más, no hay solución.

Decimos que los cuartiles dividen a la distribución ordenada de datos en cuatro partes con igual frecuencia (un 25% cada una de ellas). En sentido preciso, los cuartiles se refieren a los valores de tres posiciones exactas, no a datos concretos. Así, es correcto afirmar que Q_1 deja por debajo a 12,5 datos y por encima a 47,5, mientras que el valor 2 (pues $Q_1 = 2$) deja por debajo a 10 datos, mientras por encima se encuentran 43.

Este problema disminuye cuando tenemos muchos valores y muchos datos. En nuestro ejemplo (pensado para ser manejable a mano) no es muy relevante entrar en estas discusiones.

Los cuartiles no se utilizan para interpretar casos porque cuatro partes sigue implicando poca precisión. Los cuartiles se utilizan más para comprender el conjunto de datos, como ocurrió en la construcción de diagramas de caja y patillas que vimos en el documento *Conocer una variable*.

Para ganar en precisión, utilizamos más puntos de corte. Un ejemplo son los *deciles* (D_i): nueve cortes para conseguir diez partes con la misma frecuencia de datos. Tampoco se utilizan, salvo rara vez, para interpretar casos, sino para el mismo cometido que los cuartiles.

La medida de posición más utilizada es la que consigue cien partes con igual cantidad de datos, por lo que los 99 puntos de corte reciben el nombre de *centiles* (C_i) o también *percentiles* (P_i). Esta subdivisión es mucho más útil porque permite mejor el cometido de interpretar casos, al distinguir con más finura entre ellos. No obstante, imagina los problemas de discrepancia entre posiciones y valores que pueden tener lugar en el caso de los percentiles, si ya los hemos tenido en medianas o cuartiles. En la práctica, la solución tiene varios aspectos. Vamos a verlo.

- Hay que tener presente que el cálculo de percentiles no tiene ningún sentido con un conjunto de datos con pocos valores. Cuando una familia va a la consulta de pediatría con su bebé, los profesionales interpretan los valores de crecimiento (longitud, peso...) a partir de tablas de percentiles. Con este recurso, el profesional informa a los padres, por ejemplo, de que el niño es más alto de lo habitual porque se encuentra en el percentil 75 (es decir, es más alto que el 75% de los bebés de su sexo y edad), o que hay que vigilar su sobrepeso porque se sitúa en el P_{90}. Los percentiles se utilizan mucho para interpretar puntuaciones en test psicológicos. Según el baremo de un test de ansiedad, por ejemplo, sabemos que el nivel de ansiedad de una persona se encuentra en el percentil 28 (el 28% de la población tiene menos ansiedad que ella, o bien, el 72% tiene niveles más altos). Para conseguir esta precisión, es bueno contar con mucha riqueza de valores. En el ejemplo de pediatría, esas tablas se han construido a partir de la medida de peso y altura de muchos bebés que han generado una gran diversidad de valores.
- La interpretación de los casos según una tabla de percentiles o baremo es útil cuando el caso que se quiere interpretar cuenta con características equiparables o similares a las que se utilizaron en la muestra desde la que se generó el baremo. Pensemos en un test de agilidad mental, por ejemplo. La agilidad mengua con la edad. Imagina que utilizamos un baremo de agilidad mental generado con jóvenes de 16 a 20 años. Con bastante probabilidad, una persona de 70 años suministrará un valor que se encontrará en los percentiles más bajos. Esto es, entre otras cosas, inútil. Necesitamos un baremo de personas mayores para interpretar la agilidad mental de una persona mayor. Las niñas desarrollan sus capacidades intelectuales con más rapidez que los niños durante la pubertad. Si vamos a interpretar el desarrollo intelectual de una persona de 12 años, necesitamos utilizar el baremo construido no solo con personas de una edad similar a la suya, sino también distinguiendo entre chicos y chicas. Lo mismo ocurre en términos

culturales. Un mismo nivel de ansiedad puede ser normal en un país occidental típico, pero ser interpretado como alto en un país sureño, donde la gente valora el sosiego. La carga erótica de una imagen tiene un fuerte componente cultural. Aunque cada vez se encuentra más extendida (o impuesta) la cultura globalizada anglosajona y sus patrones son los mismos en cada vez más puntos del planeta, todavía ocurre que en diversos lugares no es raro que una mujer camine con los pechos descubiertos, mientras que en otros se considera la nuca como una zona erótica que ha de permanecer cubierta en público. Si ordenamos un conjunto de fotografías según su carga erótica, habrá que tener muy en cuenta la cultura de referencia. En psicología existe un área específica que estudia lo que se denomina *el funcionamiento diferencial de los items*. Es la constancia de que los items de un cuestionario o de un test psicológico tienen un funcionamiento muy diferente en diferentes grupos humanos, especialmente caracterizados por referentes culturales distintos.

- Por último, observa que estas medidas de posición se denominan también *percentiles*. Esto se debe a que su interpretación habitual se basa en porcentajes o tantos *por ciento*. En un ejemplo anterior he hablado de que un bebé puede tener problemas de sobrepeso al encontrarse en el P_{90}, es decir, al contar con un peso *superior al 90%* de la población de bebés de su sexo y edad. Los percentiles no tienen por qué coincidir con los tantos por ciento, pero sí parecerse mucho. Dado que las medidas de posición tienen los problemas que hemos visto en torno a la discrepancia entre posiciones y valores, y dado que podemos tener interés específico en medir posiciones desde diferentes puntos de vista, ocurre que en la práctica se utilizan fórmulas o algoritmos diversos para calcular percentiles. No todos suministran el mismo valor. Si nos ceñimos escrupulosamente a una fórmula de cálculo, entonces la utilidad de la interpretación disminuye drásticamente. Imagina a una psicóloga contando a unos padres que “su hijo muestra un nivel de extroversión que se encuentra en el percentil 87, es decir, que tiene un valor que se corresponde con la posición 193,28 que surge de ponderar la distancia equivalente entre una extroversión de valor directo 18 y otra de valor directo 18,72 que es precisamente... ”. Si te da tiempo de imaginar la cara de esos padres antes de que salgan del despacho asustados, concluirás igualmente que ese discurso no tiene sentido en ese contexto. Lo que hará la psicóloga en la práctica es decir algo parecido a “su hijo es muy extrovertido, su nivel de extroversión es superior a casi el 90% de los chicos de su edad”.

En nuestro caso, dado que los percentiles se interpretan como porcentajes, dado que vamos a manejar conjuntos de datos sin una barbaridad de valores, dado que percentiles y porcentajes acumulados coinciden aproximadamente, y dado que hemos decidido no complicarnos la vida en exceso si no hay buenas razones para ello, entonces vamos a utilizar los porcentajes acumulados como medida para los percentiles. Ten presente que esta es una decisión concreta entre un conjunto amplio de posibilidades. Si calculas los percentiles en SPSS, en LibreOffice Calc, a mano con una fórmula específica, o con cualquier otro procedimiento, no te alarmes si no coinciden con el porcentaje acumulado. No es grave. Lo importante es no perder la sensatez y saber interpretar los percentiles, se calculen como se calculen. Del mismo modo que vas a hacerlo en esta asignatura con porcentajes acumulados, así lo harás para interpretar el baremo de un test psicológico o los pesos de bebés, por ejemplo. Para ejemplificarlo, tomamos la tabla de frecuencias de nuestro ejemplo y le añadimos una fila (o columna si estuviera dispuesta en vertical): la de porcentajes acumulados.

Tabla de frecuencias																	
Xi	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
fi	3	6	7	4	3	4	5	3	3	2	2	2	2	1	1	1	1
%ai	6	18	32	40	46	54	64	70	76	80	84	88	92	94	96	98	100

Como la mayoría de los datos se encuentran agolpados en torno a los valores bajos y el conjunto de datos se dispersa conforme se aleja hacia los valores altos, puedes observar que los porcentajes altos se alcanzan con rapidez, de tal forma que el valor $X_i = 5$ muestra ya un %a = 54.

La forma en que vamos a utilizar los porcentajes acumulados como percentiles es la siguiente:

- *De percentil a valor.* Cada valor concreto ocupa un rango de porcentajes. Así, el valor 2 se asocia con el porcentaje acumulado 32, mientras que el valor 3 lo hace con %a = 40. Esto significa que los porcentajes 33 a 40 son todos *del* 3. Si nos interesa conocer el valor de percentil P_{35}, la respuesta será $P_{35} = 3$, puesto que $P_{32} < 3 \leq P_{40}$. En otras palabras, todos los percentiles con subíndice mayor de 32 e iguales o menores a 40 se corresponden con el valor 3.
- *De valor a percentil.*
  - Si el valor está presente. El percentil asociado a un valor es su porcentaje acumulado. Así para interpretar el valor 10 en la tabla, decimos que se corresponde con P_{84}, puesto que $X_i = 10 \rightarrow \%a_i = 84$.
  - Si ese valor no se encuentra explícitamente en la tabla. Se acude entonces al intervalo de valores en el que se inserta y se toma el %a del extremo menor. ¿Cuál es el percentil del valor 17? $X_i = 17$ no existe en nuestra tabla. El valor inferior más cercano es $X_i = 15$, cuyo %a = 98. Así pues, $X_i = 17 \rightarrow P_{98}$.
  - Si %a = 100. Como sabes, el porcentaje acumulado de valor 100 se encuentra en todas las tablas. Es el %a del valor máximo. Sin embargo, los percentiles son originalmente 99, puesto que se requieren *k*-1 puntos de corte para conseguir *k* partes. Luego, no existe P_{100}. Por conveniencia, hacemos $P_{100} = P_{99}$.

Para terminar este apartado vamos a resolver los dos caminos: pasar de valores a percentiles y de percentiles a valores, para todos los valores enteros entre 0 y 20 y para todos los percentiles de 1 a 99, según los datos del ejemplo.

De valores a percentiles:

Valor	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
i en P_i	6	18	32	40	46	54	64	70	76	80	84	88	92	94	96	98	98	98	98	98	99

De percentiles a valores:

(El subíndice del percentil se divide entre unidad y decena. Las unidades se encuentran en columnas. Las decenas en filas. Para encontrar el valor del percentil P_{38}, dado que 38 =30+8, mira la columna 8 y la fila 30, con lo que encuentras que P_{38}=3)

	0	1	2	3	4	5	6	7	8	9
0		0	0	0	0	0	0	1	1	1
10	1	1	1	1	1	1	1	1	1	2
20	2	2	2	2	2	2	2	2	2	2
30	2	2	2	3	3	3	3	3	3	3
40	3	4	4	4	4	4	4	5	5	5
50	5	5	5	5	5	6	6	6	6	6
60	6	6	6	6	6	7	7	7	7	7
70	7	8	8	8	8	8	8	9	9	9
80	9	10	10	10	10	11	11	11	11	12
90	12	12	12	13	13	14	14	15	15	20

Distancias estandarizadas

Las medidas de posición se utilizan con mucha frecuencia, pero existe otro recurso más extendido todavía: las distancias estandarizadas. Cuentan con cinco importantes ventajas frente a los percentiles:

1. No sufren los problemas de discrepancia entre posiciones y valores.
2. Son fáciles de calcular en términos aritméticos. Observa que los percentiles surgen de poner en práctica un algoritmo no aritmético (busca la posición, identifica el valor), mientras que en estadística existe una comprensible predilección por las herramientas que surgen íntegramente del cálculo, de tal forma que, por ejemplo, son fáciles de obtener en una calculadora, una hoja de cálculo o un programa aritmético de ordenador. Además, las operaciones aritméticas permiten realizar deducciones con relativa comodidad.
3. Se utilizan como resultado o producto intermedio en varias herramientas estadísticas muy frecuentes en el análisis de los datos.
4. Los percentiles no aprovechan toda la información de una variable cuantitativa, sino que utilizan únicamente su orden. No obstante, las distancias estandarizadas consideran toda la información original de la variables porque también tienen en cuenta la cuantía.
5. Las distancias estandarizadas admiten transformaciones aritméticas, lo que permite generar variables muy útiles para facilitar la interpretación psicológica, como veremos más adelante.

A pesar de estas ventajas y como vamos a ver, las distancias estandarizadas cuentan con un inconveniente serio respecto a los percentiles: son más difíciles de interpretar. Lo son por dos razones: para comprenderlas es necesario poner en juego más conocimiento estadístico. Y también porque generan resultados con decimales y signo. Si una persona lee el valor 9, no pasa nada. Pero si lee -0,013, es posible que le resulte desagradable. Así somos. La mayoría de las distancias estandarizadas en la mayoría de los conjuntos de datos se encuentran entre -2 y +2. A la hora de comunicarse con personas (sean o no profesionales de la psicología) los percentiles son más recomendables que las distancias estandarizadas porque no hay que dar un curso de estadística para que nuestro interlocutor entienda lo que le estamos diciendo. Aún así, algunos baremos en psicología utilizan distancias estandarizadas *transformadas*. Veremos esto más adelante, en el subapartado *transformación de escala*.

Estandarizar una distancia

La idea fundamental de una distancia estandarizada, como puede deducirse de su nombre, es calcular la distancia directa que existe entre un valor concreto y un valor que se utiliza como referente. Acto seguido, esa distancia sufre una transformación que denominamos *estandarización* y cuyo objetivo es facilitar la interpretación y la comparabilidad entre varias distancias. Esta comparabilidad es posible porque la estandarización es sensible a las características del conjunto de datos. De esta forma, aunque dos valores provengan de dos conjuntos diferentes de datos (por ejemplo, dos muestras de personas de culturas diferentes, o resultados de exámenes en dos universidades), podemos interpretar y comparar casos concretos si sus valores se han estandarizado.

Lo primero, pues, es escoger ese *valor de referencia* que utilizamos para calcular las distancias. Se trata de la media aritmética. Es la mejor opción, desde el conocimiento que tenemos de la estadística en este curso. Es la mejor opción porque el cálculo de distancias, y más aún su estandarización, requiere de los datos que sigan una escala cuantitativa. Como sabemos, de las representaciones numéricas o medidas de tendencia central, la inicialmente más recomendable para variables cuantitativas es la media aritmética. Así que la escogemos como valor representativo del conjunto de datos.

La primera fase del cálculo de distancias estandarizadas es obtener la distancia que separa a cada valor respecto a la media de su conjunto de datos, es decir:

$$X_i - \bar{X}$$

Esta operación suministra ya una información muy útil: si el resultado es negativo, sabemos que el valor se encuentra por debajo de la media aritmética; en caso contrario, se encuentra por encima; y si el valor es cero, entonces coincide con la media. Una diferencia de -7, por ejemplo, indica que el valor original se encuentra a 7 puntos de la media y que es inferior a esta. Es una información útil pero muy incompleta. No sabemos interpretar la cuantía de la distancia, es decir, si 7 es mucha diferencia o poca. Para esa interpretación podríamos utilizar un criterio absoluto, como por ejemplo “más de 5 puntos de distancia en la variable *introversión* implica la existencia de serios problemas de comunicación con los demás”. En la práctica, estos criterios absolutos son una rareza. No solemos contar con ellos. Así que utilizamos un mecanismo relativo: comparamos la distancia que queremos interpretar en el conjunto de distancias. En otras palabras, si los datos del conjunto de referencia están muy distantes entre sí y, por lo tanto, también de la media, entonces la distancia que nos preocupa será mucho menos importante que si los datos están muy cercanos a la media. Así, si lo habitual es alejarse en torno a 10 unidades, una distancia de 7 es poca cosa. Pero si lo habitual es alejarse 3, una distancia de 7 es muy importante.

Hemos visto un recurso para medir cómo de distantes se encuentran los datos respecto a su media aritmética: la desviación tipo. Recuerda:

- Una buena idea para medir cuán distantes están los datos entre sí es calcular la media de las distancias a la media, es decir, cuánto de distantes están los datos por término medio.
- Pero la suma de las distancias a la media siempre vale cero, lo que es lógico porque a ambos lados de la media contamos con la misma suma de distancias.
- Así que se elevan las distancias al cuadrado para evitar que las distancias negativas se anulen con las positivas.

– La media de las distancias al cuadrado se denomina varianza. Pero tiene un inconveniente: se expresa en unidades al cuadrado (por ejemplo, la varianza de la variable edad puede ser 78 años cuadrados ¿qué es un año al cuadrado?).
– Así que nos inventamos la desviación tipo: la raíz cuadrada de la varianza.

Para interpretar cuán importante es una distancia calculada en un conjunto de datos, podemos expresarla utilizando la desviación tipo de ese conjunto como unidad de medida. En otras palabras: la distancia se divide por el valor de la desviación tipo. Yo solía hacer eso con los bocadillos. Cuando estudiaba psicología era habitual que almorzara un bocadillo. Cuando alguien me decía cuánto valía algo (un coche, una casa, unos pantalones...), yo dividía su coste por lo que me costaba a mí un bocadillo. Con ese recurso, expresaba el valor de ese objeto en el número de bocadillos a que equivalía, una medida para mí con mucho sentido y que resiste el paso del tiempo. Si me mudo a otro punto del planeta y sigo almorzando bocadillos, seguirá siendo un buen recurso para interpretar el precio de las cosas. Así, por ejemplo, si el bocadillo que suelo tomarme cuesta 2 euros, y los zapatos que dices que me compre cuestan 70, yo pienso “por ese precio almuerzo yo 70/2=35 días”. Mi unidad de medida es el bocadillo.

Con las distancias ocurre lo mismo. Dado que cada conjunto de datos tiene su propia media (que expresa sobre qué valor ronda el conjunto) y su propia desviación tipo (que expresa cuán distantes se encuentran los datos respecto a la media), una buena manera de interpretar un caso concreto es re-expresarlo como el número de desviaciones tipo que se encuentra alejado de la media. Por eso lo llamamos *distancia estandarizada*: se estandariza. Esto permite comparar datos que provienen de conjuntos distintos. Imagina que vas de vacaciones a un país donde no has estado nunca. Una de tus preocupaciones es el dinero que debes llevar para no tener problemas. Si utilizas el referente del país donde vives puedes llevarte una sorpresa (agradable o desagradable). Para interpretar cuán cara o barata es la vida a donde vas, un buen criterio es averiguar cuánto cuestan las cosas que sueles consumir, como la comida o los medios de transporte. Tienes que situarte en ese lugar para interpretarlo, no te servirá utilizar el tuyo.

¿Qué significa que hoy ha hecho mucho calor? ¿Cuántos grados centígrados es eso? La respuesta es “depende”. 30 grados es una temperatura muy agradable en el verano andaluz, pero muy caluroso en el verano danés. Depende del conjunto de datos en donde se inserte, un valor concreto se interpretará de formas muy diferentes.

Las distancias estandarizadas reciben varias denominaciones. En términos habituales, se las conoce mejor como *puntuación*. Ese es el nombre. El apellido es el mismo que la desviación tipo, también con varios apellidos: tipo, típica o estándar. Por este motivo, nuestra distancia estandarizada también se conoce como puntuación tipo, puntuación típica o puntuación estándar. Y suele denotarse con el símbolo Z.

$$Z_i = \frac{X_i - \bar{X}}{S}$$

Nuestro ejemplo

Vamos a observar cómo son las distancias estandarizadas para nuestro ejemplo.

Tabla de frecuencias																	
Xi	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
fi	3	6	7	4	3	4	5	3	3	2	2	2	2	1	1	1	1
%ai	6	18	32	40	46	54	64	70	76	80	84	88	92	94	96	98	100
Zi	-1,28	-1,05	-0,83	-0,61	-0,38	-0,16	0,07	0,29	0,52	0,74	0,96	1,19	1,41	1,64	1,86	2,09	3,21

Los valores más raros o peculiares, en el sentido de ser especialmente distantes, se encuentran en la parte de las distancias positivas o valores altos en el caso de nuestro ejemplo (en otros ejemplos, pasará lo contrario). El valor $X_i = 20$ se encuentra a 3,21 desviaciones tipo de la media. Teniendo en cuenta que suele ser infrecuente encontrar distancias superiores a 2, este valor estandarizado expresa una distancia muy sobresaliente. El siguiente dato muy alejado es el correspondiente al valor $X_i = 15$ ($Z_i = 2,09$). Observa que las distancias negativas son relativamente pequeñas en nuestro ejemplo. La distancia negativa (valor por debajo de la media del conjunto) más sobresaliente es $Z_i = -1,28$ que se corresponde con $X_i = 0$. Hay 6 valores en el extremo superior con valores más alejados (desde $X_i = 12$ hasta $X_i = 20$, ambos inclusive). Este comportamiento sigue expresando lo que ya sabemos: los datos están agolpados en el lado de los valores bajos y se dispersan hacia los altos.

Propiedades interesantes

Las distancias estandarizadas o puntuaciones tipo tienen unas cuantas propiedades matemáticas muy interesantes. La principal de todas ellas la hemos abordado ya, aunque con otras palabras: es independiente de la escala utilizada. Aunque conceptualmente ya sabemos que Z consigue expresarse en una unidad (número de desviaciones tipo) que permite interpretar casos aunque provengan de conjuntos distintos de datos, vamos a verlo de otras formas. Lo primero es un ejemplo.

Imagina que estamos utilizando el metro como unidad para expresar las longitudes de las calles de tu barrio. Tu calle mide 85 metros. ¿Es larga, corta o normal? Si la media de longitudes es 85, ocurre que la tuya representa muy bien las dimensiones de las calles de tu barrio. Si la media es 100, tu calle es corta, aunque no sabemos si mucho o poco. Si S = 5, entonces Z = -3 y tu calle es muy corta. Si S = 30, entonces Z = -0,5 y tu calle es levemente corta, prácticamente normal.

Pongamos que $X_i = 85$, la media es 100 y la desviación tipo 10. Entonces:

$$Z_i = \frac{X_i - \bar{X}}{S} = \frac{85 - 100}{10} = -1,5$$

La distancia estandarizada que representa la longitud de tu calle es -1,5 metros. Esto significa que tu calle es claramente más corta que la media. ¿Y si medimos en centímetros? Al cambiar la unidad de medida todo queda afectado. Tu calle ya no mide 85, en metros, sino 8500, en centímetros. La media aritmética ya no son 100 metros sino 10 mil centímetros. La desviación tipo ya no tiene el valor de 10 metros sino de 1000 centímetros. ¿Y tu Z? Veamos:

$$Z_i = \frac{X_i - \bar{X}}{S} = \frac{8500 - 10000}{1000} = -1,5$$

¡La puntuación tipo es la misma! No es sorprendente, pues se trata de una distancia *estandarizada*, es decir, independiente de la escala. Este comportamiento es la

principal fuerza de Z, pues permite comparar datos aunque provengan de conjuntos donde se han utilizado escalas diferentes. Veamos un ejemplo de comparación.

Belén ha obtenido un 6 en el examen de geografía. Susana ha obtenido un 7. Si el examen está bien hecho, deberíamos concluir que Susana sabe más que Belén en asuntos de geografía. No obstante, cada una asiste a clase en un instituto de enseñanzas medias diferente. Los exámenes han sido también diferentes. Eso hace que no podamos comparar las calificaciones de ambas. Es posible que la prueba de Susana sea más sencilla y, por tanto, sea más fácil obtener calificaciones altas que en el caso de la prueba a la que se ha enfrentado Belén. Pero... un momento... Si el examen es más difícil, lo será presumiblemente para toda la clase de Belén y la nota media de la clase de Susana debería ser más elevada. Si restamos la media a ambas calificaciones podremos eliminar la influencia de la dificultad. Aunque...

Hay otro problema. Es posible que los exámenes no se midan ambos en la escala de 0 a 10, o que la variabilidad de resultados sea mayor en una clase que en la otra. Tal vez cada problema tenga un peso de un punto en ambos casos, pero el número de problemas sea diferente. Para controlar la escala, que a su vez controla la calificación media, entonces estandarizamos ambos resultados y ello nos permitirá resolver la duda de la comparación. Vamos a imaginar varias situaciones, ya que me estoy inventando el ejemplo y ello me da la libertad de hacer lo que me dé la gana. Así que vamos a aprovecharlo para partir de situaciones diferentes que, manteniendo la mismas calificaciones, nos llevan a conclusiones diferentes.

		Situación A			Situación B			Situación C		
	Xi	Media	Des.tip.	Z	Media	Des.tip.	Z	Media	Des.tip.	Z
Belén	6	5	1	1,00	5	1	1,00	5	0,5	2,00
Susana	7	5	1	2,00	5	2	1,00	8	0,5	-2,00

En la situación A, las medias y las desviaciones tipo de ambas pruebas coinciden, por lo que no añaden nada a la comparación y basta con conocer las puntuaciones directas u originales. La conclusión es que, según esta información (las situaciones reales son siempre más ricas o complejas), Susana controla mejor la geografía que Belén. Es más, Susana sabe bastante, puesto que el valor de su puntuación estándar es muy pronunciado (ya he dicho que Z=2 es una distancia poco habitual).

En la situación B, las medias de ambas pruebas coinciden, pero la dispersión de las calificaciones es más elevada en la clase de Susana que en la de Belén. Al final, ambas han obtenido una calificación que supera a la media de su clase en 1 desviación tipo, por lo que ambas tienen la misma calificación estándar y sus conocimientos en geografía son, por tanto, equivalentes.

Por último, en la situación C las cosas cambian bastante. Lo primero que destaca es que parece que el examen que ha hecho Susana es más sencillo que el realizado por Belén (si la preparación del conjunto de la clase es equivalente, claro), por lo que la gente que acompaña a Susana ha elevado sensiblemente la media. Ocurre entonces que Susana ha obtenido una calificación inferior a la media de su clase, mientras que a Belén le ha ocurrido lo contrario. No acaba ahí la cosa. Ambas clases muestran una dispersión pequeña, pues S = 0,5 traducido a CV de Pearson es de 10% para el grupo de Belén y de 6,25% para el grupo de Susana, dos valores que implican dispersiones muy bajas. Con poca dispersión, una misma distancia es más importante. Por eso, aunque las notas siguen siendo 6 y 7, su importancia ahora es mayor, pues están muy separadas de la media: Belén supera el valor medio de su clase en dos desviaciones tipo, mientras que a Susana le ocurre lo contrario, queda por debajo del valor característico de su clase en dos

desviaciones tipo. Ambas han obtenido calificaciones muy diferentes y aunque Susana tiene una puntuación directa superior a la de Belén, controla mucho menos la geografía.

Además de esta capacidad para controlar la escala y permitir que comparemos puntuaciones entre sí, hay otras peculiaridades aritméticas interesantes: la media de las puntuaciones tipo siempre es 0 y su desviación tipo siempre es 1.

Recuerda que la media de las distancias a la media siempre es 0. De ahí se deduce que la media de las puntuaciones tipo (que son distancias a la media, estandarizadas) también es 0. En el anexo a este documento tienes las demostraciones para estas afirmaciones. La demostración de que la media de las puntuaciones tipo siempre vale 0 es la número 7. La 9 es la demostración de que la desviación tipo de las puntuaciones tipo siempre vale 1.

Transformación de escala

Las dos propiedades que acabamos de ver (media 0 y desviación tipo 1) tienen más consecuencias de las que tal vez puedas pensar ahora mismo. Vamos a ver una en concreto con muchas repercusiones, especialmente en psicología. Ocurre que cuando sumamos una constante a un conjunto de números, la media se ve afectada del mismo modo. Si sumamos, por ejemplo, el valor 5 a todos los datos, entonces la media resultante también se ve incrementada en 5 unidades. Gráficamente es como si desplazamos todo el conjunto de datos 5 unidades hacia la derecha, lo que afecta igualmente a la media, pero deja a la dispersión exactamente igual (la gráfica sigue siendo igual de ancha o estrecha). Si lo que hacemos no es sumar sino multiplicar, la media y la desviación tipo también se ven multiplicadas por el mismo valor.

Pues bien, si multiplicamos las puntuaciones tipo por una constante de valor *a*, su desviación tipo y su media se ven también multiplicadas por *a*. Dado que $S_Z = 1$ entonces, la nueva desviación será exactamente *a*. Y dado que la media de Z vale 0, la media seguirá valiendo 0 (pues 0 por cualquier valor sigue siendo 0). Y si a todos los datos le sumamos una constante *b*, entonces, la nueva media valdrá exactamente *b* (pues 0+b=b) y la desviación tipo no se verá afectada (por mucho que sumemos, las distancias no varían). En otras palabras, si queremos generar una variable que tenga de media exactamente *b* y de desviación tipo exactamente *a*, entonces la nueva variable surge de hacer la siguiente transformación a partir de las puntuaciones tipo de la variable original:

$$V_i = aZ_i + b$$

La demostración 11 justifica que la nueva variable tendrá de media el valor *b*, mientras que la demostración 14 hace lo mismo con la desviacion tipo de valor *a*.

Estas operaciones son útiles en psicología porque podemos construir un baremo de puntuaciones transformadas con la media y la desviación tipo que deseemos. Por ejemplo, en los test de cociente intelectual, primero se obtienen las puntuaciones directas, que pueden ir de 0 a 200 puntos (no te lo tomes literalmente, me lo acabo de inventar). Después se tipifican o estandarizan, con lo que las puntuaciones tipo resultantes sabemos que tienen 0 de media y 1 de desviación tipo. Son puntuaciones de valores bajos (recuerda que es poco habitual alejarse más de 2 desviaciones tipo de la media) que además manejan decimales y donde buena parte de los valores son negativos. Estas tres características son incómodas para el común de los mortales. Así que realizamos dos transformaciones:

1. Se multiplican las puntuaciones tipo por un número, por ejemplo 20, que permita prescindir de los decimales porque amplía la escala de medida. Tras la multiplicación, las puntuaciones resultantes irán de -40 a +40 en su mayoría, una amplitud suficiente como para hacer prescindible la coma decimal.

2. Se suma una constante que haga muy difícil llegar a valores negativos y que tenga un sentido psicológico positivo. Si sumamos, por ejemplo, el valor 100, la mayoría de los datos se encontrarán entre 60 y 140, con media en 100, un valor que manejamos con cierta comodidad.

Cuando revises la documentación de un test psicológico encontrarás los baremos expresados en puntuaciones centiles o percentiles, o bien en puntuaciones típicas transformadas de un modo similar al que hemos visto aquí.

Si aplicamos la transformación $V_i = 10Z_i + 50$ a los datos del ejemplo, obtenemos:

Tabla de puntuaciones tipo transformadas

Xi	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
Zi	-1,28	-1,05	-0,83	-0,61	-0,38	-0,16	0,07	0,29	0,52	0,74	0,96	1,19	1,41	1,64	1,86	2,09	3,21
Vi	37	39	42	44	46	48	51	53	55	57	60	62	64	66	69	71	82

Debido a la transformación que hemos realizado, la mayoría de los datos se van a encontrar entre 30 y 70 (dos desviaciones tipo por encima y por debajo de la media). Observa que hay dos casos que se salen de esa norma, los que corresponden a $X_i = 15$ y muy especialmente $X_i = 20$. La escala V_i es cómoda y fácil de interpretar.

Anexo: demostraciones

1. $$\sum X_i = \frac{n}{n}\sum X_i = n\frac{\sum X_i}{n} = n\bar{X}$$

2. $$\sum \bar{X} = \sum_{i=1}^{n} \bar{X} = \bar{X}\sum_{i=1}^{n} 1 = n\bar{X}$$

3. $$\sum \left(X_i - \bar{X}\right) = \sum X_i - \sum \bar{X} = (1\ y\ 2) = n\bar{X} - n\bar{X} = 0$$

4. $$\sum Z_i = \frac{\sum X_i - \bar{X}}{S} = \frac{\sum \left(X_i - \bar{X}\right)}{S} = (3) = \frac{0}{S} = 0$$

5. $$\sum \left(X_i - \bar{X}\right)^2 = \frac{n}{n}\sum \left(X_i - \bar{X}\right)^2 = n\frac{\sum \left(X_i - \bar{X}\right)^2}{n} = nS^2$$

6. $$\sum Z_i^2 = \sum \left(\frac{X_i - \bar{X}}{S}\right)^2 = \frac{\sum \left(X_i - \bar{X}\right)^2}{S^2} = (5) = \frac{nS^2}{S^2} = n$$

7. $$\bar{Z} = \frac{\sum Z_i}{n} = (4) = \frac{0}{n} = 0$$

8. $S_z^2 = \dfrac{\sum \left(Z_i - \bar{Z}\right)^2}{n} = (7) = \dfrac{\sum Z_i^2}{n} = (6) = \dfrac{n}{n} = 1$

9. $S_z = \sqrt{S_z^2} = (8) = \sqrt{1} = 1$

Con $V_i = aZ_i + b$ entonces:

10. $\sum V_i = \sum \left(a Z_i + b\right) = a \sum Z_i + \sum b = (4) = \sum b = nb$

11. $\bar{V} = \dfrac{\sum V_i}{n} = (10) = \dfrac{nb}{n} = b$

12. $$\sum V_i^2 = \sum \left(a Z_i + b\right)^2 = \sum \left(a^2 Z_i^2 + b^2 + 2\text{ab}Z_i\right) =$$

$$= a^2 \sum Z_i^2 + \sum b^2 + 2\text{ab} \sum Z_i = (4 \text{ y } 6) = a^2 n + n b^2 + 0 = n\left(a^2 + b^2\right)$$

$$S_V^2 = \frac{\sum \left(V_i - \bar{V}\right)^2}{n} = (11) = \frac{\sum \left(V_i - b\right)^2}{n} = \frac{\sum \left(V_i^2 + b^2 - 2\text{b}V_i\right)}{n} =$$

13. $$\frac{\sum V_i^2 + \sum b^2 - 2\text{b} \sum V_i}{n} = (10 \text{ y } 12) = \frac{n\left(a^2 + b^2\right) + nb^2 - 2\text{nb}^2}{n} =$$

$$= a^2 + b^2 + b^2 - 2\text{b}^2 = a^2$$

14. $S_V = \sqrt{S_V^2} = (13) = \sqrt{a^2} = a$