Conocer una variable

Vicente Manzano-Arrondo – 2010,2013-2014

Aunque no puede entenderse un dato sin conocer de dónde viene y hacia dónde va, lo primero específicamente estadístico que se requiere hacer ante cualquier conjunto de datos es llevar a cabo un análisis variable a variable. Es necesario conocer a cada una, cómo se comporta, si hay errores de transcripción, si existen casos raros o anómalos que pudieran distorsionar las conclusiones generales y que aconsejan un estudio específico, si hay ausencias de información que requieren un proceso de imputación o decisión de valores... Sin conocer a cada variable por separado, iniciaríamos los análisis con una notable desventaja y una probable gestación de conclusiones erróneas, sesgadas o incompletas.

Para llevar a cabo un estudio de cada variable es importante conocer primero de qué tipo es según el procedimiento de medida que se ha llevado a cabo. En este documento vamos a distinguir cuatro clases: nominal, ordinal, cuasicuantitativa y cuantitativa.

Tomar buenas decisiones respecto a qué hacer con una variable en concreto depende no sólo del procedimiento de medida sino también de otras circunstancias. No es recomendable generar un amplio listado de consejos para generar buenas decisiones. Lo que vamos a hacer es recurrir a una doble dimensión. La primera es la normativa, o conjunto de criterios sobre cómo utilizar un instrumento. Por ejemplo, diremos que el instrumento “ciclograma” es la mejor opción para una variable nominal. La segunda dimensión es la del sentido común y el dominio de los significados. Así, aunque un ciclograma es una buena opción en términos generales para una variable nominal, no olvidemos que el objetivo de una representación gráfica es expresar con rapidez, claridad y sin error, sesgo o engaño, las características principales del comportamiento de una variable. Si al realizar el ciclograma observas que no se cumple esta función, habrá que buscar otra alternativa. Puede ocurrir, por ejemplo, que se hayan manejado tantas categorías en la variable, que el ciclograma quede precioso en términos estéticos, pero inútil en términos estadísticos, pues sirve al objetivo de adornar mi habitación como un póster, pero no para expresar de qué va la variable. No esperes, en este caso, un consejo del tipo “si la variable nominal posee más de diez categorías, utilícese un diagrama de barras”. La respuesta a la pregunta “¿cuántas categorías marcan el máximo para que un ciclograma deje de ser una buena opción?”, es “depende”, una contestación muy frecuente en este campo de conocimiento. Utiliza los significados y el sentido común, no olvides cómo se comporta cada instrumento de análisis ni los objetivos específicos de cada uno de tus movimientos en estadística. Puede ocurrir, por ejemplo, que la variable nominal cuente con cien categorías, pero 97 de ellas tienen una frecuencia mínima, por lo que el ciclograma mostrará una información muy clara y fácilmente interpretable, salvo en un sector donde se amontonan 97 categorías que pueden percibirse perfectamente como del tipo “otros”. A lo largo de este documento intentaré mostrar con claridad la norma y entrar en algunas excepciones para ejercitar el uso de los significados. Si el sentido común pudiera normativizarse, ya no sería sentido común.

Dentro de cada tipo de variable vamos a observar su definición, su tabulación, su representación gráfica y su representación numérica, además de algunas matizaciones.

Variable nominal

Como ya conoces, las operaciones aritméticas que se pueden realizar con los datos numéricos de una variable dependen del procedimiento que se ha seguido para asignar números a los diferentes acontecimientos considerados de esa variable. Dado que el análisis de datos implica operaciones aritméticas, es fundamental saber qué cosas se pueden hacer y cuáles no con un tipo concreto de variable.

Escala

En una nominal, los números expresan sencillamente identidades. Dos números diferentes indican que se refieren a dos estados diferentes de la variable. Y nada más. Si hemos preguntado por el color del cabello, es posible haber codificado “moreno” con el valor 1, “rubio” con el 2, “castaño” con el 3 y “otros” con el 4. Dado que el número 4 es diferente al 3, está expresando que apunta a un color de cabello diferente al que apunta el número 3, pero nada más. Aunque consideramos que el número 4 es mayor o superior al 3, no ocurre así en una variable nominal. Observa que “otros” no es más color que “castaño”. Puede ser más difuso, amplio o común, pero no más color de cabello. Así que en este caso, 4 no es más que 3, sino únicamente distinto.

Tabulación

Lo primero que se requiere hacer en una variable es su tabulación, es decir, organización de los datos de tal forma que resulte más sencillo comenzar a conocer la situación que no recurriendo a la matriz original. El recurso es denominado *tabla de frecuencias*, puesto que es una tabla donde se disponen todos los valores observados y sus correspondientes frecuencias. El concepto estadístico *frecuencia* coincide con el popular: número de veces que ocurre algo. Observa la siguiente tabla.

X_i	f_i
1	15
2	20
3	10
4	5
Σ	50

La primera columna es la de valores. Se representa con Xi utilizando X como letra para expresar a la variable “color del cabello”. El símbolo X_1 representa el valor 1 o moreno. Del mismo modo, X_3=castaño. En términos generales, utilizamos X_i, donde i es cualquiera de las posiciones existentes, desde la 1 a la 4. El símbolo Σ es la letra griega *sigma* que en análisis de datos significa *suma* o *sumatorio*. Se refiere al total de personas cuyo color de cabello hemos registrado. Observa que han sido 50. La segunda columna es la de frecuencias. Observa que f_2=20, por ejemplo. Lo que significa que hay 20 personas en quienes hemos observado el color de cabello rubio. Como es de sentido común, al sumar todas las frecuencias, es decir, cuántas personas tienen cada uno de

todos los colores presentes, lo que obtenemos es el número total de personas observadas. Suele utilizarse la letra *n* para expresarlo, por lo que podemos escribir *n*=50.

Lo que acabamos de ver es la versión mínima de una tabla de frecuencias. Ya que hemos organizado los datos en la tabla, es fácil añadir más información, esta vez con algo de proceso previo. La tabla es una oportunidad para expresar más cosas que nos resulten útiles para conocer la variable. Por ejemplo, las frecuencias tal y como se encuentran requieren dos informaciones para ser comprendidas. Imagina que hablas con alguien por teléfono y le comunicas cuántas personas tienen el color de cabello castaño. ¿Qué le dirías? Podría ser: “he observado diez personas con pelo castaño”. Es una información insuficiente, puesto que el receptor de esa noticia no puede hacerse una idea de la envergadura o del significado de esa cantidad. Necesita conocer cuántas personas has observado, de tal forma que la frase quedaría mejor así: “he observado que 10 de 50 personas tienen el color de pelo castaño”. No es lo mismo 10 de 50 (una quinta parte), que 10 de 10 (todas las personas observadas) o 10 de 1000 (la centésima parte), por ejemplo.

En lugar de tener que expresar necesariamente la frecuencia observada y el total, existe un recurso muy extendido mediante el que utilizamos un único número: el tanto por ciento o porcentaje. Observa cómo queda la tabla ahora.

X_i	f_i	$\%_i$
1	15	30
2	20	40
3	10	20
4	5	10
Σ	50	100

En lugar de afirmar “10 de 50”, decimos “20 por ciento”. En esencia es lo mismo: en lugar de utilizar el referente 50 acudimos al referente 100. Es algo así como “si lo que yo hubiera considerado fuera 100 personas, entonces habría observado a 20 con el cabello castaño”. La diferencia entre 50 y 100 es que este segundo caso se utiliza con tanta asiduidad que resulta muy familiar y requiere un mínimo de esfuerzo mental para ser comprendido. Por esta razón los medios de comunicación recurren a los porcentajes para exponer resultados de estudios. Observa que la suma de todos los porcentajes da como resultado $\Sigma\%_i=100$. Esto no es para memorizar. Resulta de sentido común, puesto que la suma de todo tiene que suministrar el todo y este es el 100%.

La forma de calcular un porcentaje es fácil y solemos hacerlo sin excesiva dificultad:

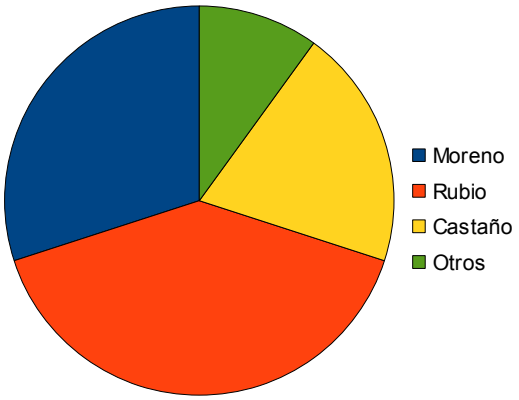
$$\frac{f_i}{n}=\frac{\%_i}{100} \quad \rightarrow \quad \%_i=\frac{100\,f_i}{n}$$

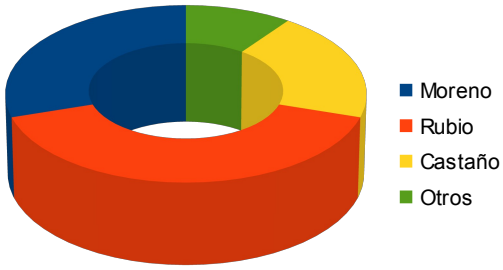
Representación gráfica

Por lo general, las variables nominales suelen contar con pocas categorías, por lo que una tabla de frecuencias es suficiente para tener una idea aceptable sobre lo que está ocurriendo con esa variable. No obstante, es posible preferir una representación gráfica a la tabla o utilizar esta para nuestro conocimiento y la gráfica para comunicarnos con otras personas a quienes vamos a exponer qué ocurre con esa variable.

A diferencia de una tabla, en la representación gráfica no se busca tanto la precisión como la exposición rápida, clara y sin errores. De un solo vistazo, quien ve la representación debe ser capaz de conocer qué ocurre con esa variable.

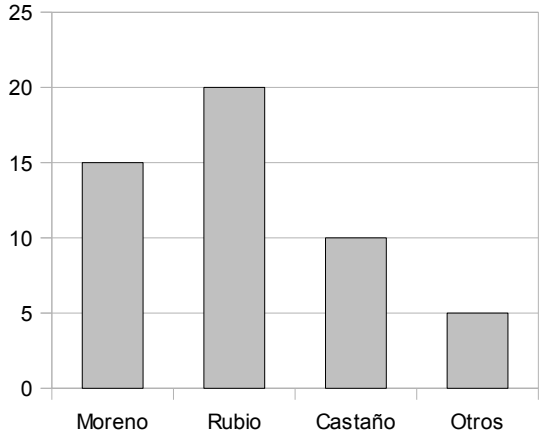
Dado que una variable nominal sólo puede manejar identidades, no órdenes ni cuantías, la representación gráfica debe ser acorde con ello. Una buena opción es el ciclograma, también llamado diagrama de sectores o gráfico de pastel. Consiste en un círculo dividido en sectores o *quesitos*. Cada sector expresa a una de las categorías observadas. El sector es tanto mayor (su ángulo es más abierto) cuanto mayor sea el valor de la frecuencia que representa. En el caso del ejemplo sobre colores de cabello, el ciclograma es éste:



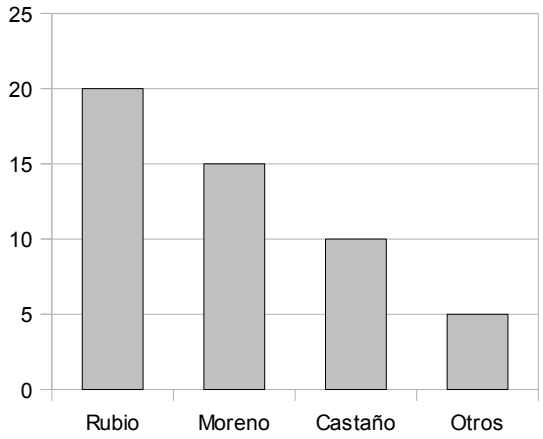


Recuerda: el objetivo es expresar con rapidez, claridad y sin error... En la construcción de gráficas sufrimos una fuerte tentación: conseguir algo especialmente bonito o impresionante. Uno de los recursos para conseguirlo es acudir a efectos 3D, separar sectores, generar figuras... Observa a la izquierda de este párrafo los mismos datos pero en otro modelo. Para la mayoría de los gustos, el resultado es más agradable a la vista. Pero para interpretar, es más difícil. La perspectiva, por ejemplo, provoca que el sector situado más cerca de quien recibe la imagen se observe con mayor importancia. Para construir un pequeño imán de nevera, un pin del chaleco, un dibujo para la carpeta o un póster para la habitación, esta segunda opción es mejor. Para interpretar datos, indudablemente cuando más sencilla sea la representación, cuantos menos recursos y adornos posea, mucho mejor.

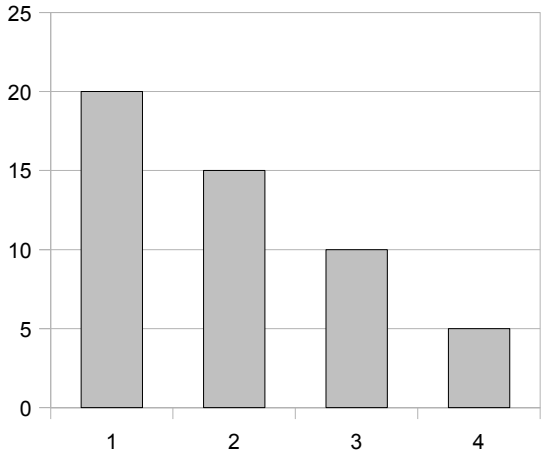
El número de grados de un ángulo requiere cierto esfuerzo mental. Nos resulta más fácil interpretar la altura o la longitud, que no la apertura de un ángulo. Por esta razón, otra opción aceptable para representar variables nominales es lo que se denomina *diagrama de barras*. Consiste en un sistema de dos ejes perpendiculares. En el eje horizontal o de abcisas se sitúan equidistantes los valores de la variable. En el eje vertical o de ordenadas se representan las frecuencias (no importa si son frecuencias observadas en valor absoluto o bien porcentajes, puesto que el resultado es el mismo). Sobre cada punto del eje horizontal se alza una barra de igual anchura, pero con una altura variable: la altura representa la frecuencia, de tal forma que conforme más arriba llegue una barra, mayor es el número de casos que se han observado para ese valor del eje horizontal. Observa cómo queda un diagrama de barras para la variable “color del cabello”.



Por sencillez, he optado por representar todas las barras con el mismo color y de tal modo que éste no sea llamativo. Observa cómo es rápido y sencillo interpretar la importancia de cada categoría o valor. El color rubio se observa rápidamente como el más frecuente, seguido de moreno, castaño y otros. El diagrama de barras es una buena opción. Sin embargo, observa ahora la siguiente representación.



Son los mismos datos, pero ahora el orden de las categorías es distinto: he puesto el color moreno en segundo lugar en vez del primero. La interpretación debería ser exactamente la misma. Y nada se opone a cambiar el orden, puesto que estamos tratando una variable nominal y la disposición de los valores es arbitraria. Observa que casi es inevitable sentenciar algo parecido a “conforme se avanza en el color hacia la derecha, disminuye la frecuencia”. Si en lugar de los valores literales utilizamos los numéricos (cambiando la codificación original, de tal forma que Rubio=1 y Moreno=2), obtenemos esto:



Fíjate en el efecto. Tenderíamos a concluir que conforme aumenta el valor de la variable, disminuye la frecuencia. Pero esto no es cierto. El valor de la variable no es numérico, sino que recurre a categorías sin orden. Cualquier disposición de las categorías en el eje horizontal es igualmente aceptable. Este efecto ocurre porque tendemos a interpretar el diagrama de barras como si el orden fuera relevante. Esto no ocurre con el ciclograma. Por esta razón, a pesar de que el ángulo de los sectores requiere mayor esfuerzo para ser interpretado, evita la tendencia que sufrimos de utilizar el orden, por lo que podemos considerar que en las variables nominales el ciclograma es mejor opción que el diagrama de barras, si bien este es también aceptable. Digamos que no es mala estrategia generar ambos gráficos y decidir, en función del resultado, cuál cumple mejor la función en cada situación concreta.

Representación numérica

Alguien puede preguntar "¿Cuál es el color más característico entre los observados?". El objetivo puede ser regalar un tinte de pelo, o preparar una fotografía del grupo, donde el conocimiento sobre los colores presentes puede resultar fundamental para el artista. Sea cual fuere el motivo, nos encontramos con la frecuente necesidad de tomar un único valor como representación de todos los valores presentes.

Existen muchos criterios a los que podríamos acudir para tomar una buena decisión. En una variable nominal, como es el caso, una buena estrategia es escoger el valor más frecuente. En este caso es la categoría "Rubio". Dado que es *el valor que más se lleva*, lo denominamos del mismo modo que hacemos en la vida cotidiana: moda. Por tanto, la moda (Mo) es el valor con mayor frecuencia.

En análisis de datos, como ocurre en psicología, procuramos conocer la norma, lo general, al mismo tiempo que tener presente lo particular, el caso concreto. En esta dinámica, es fundamental manejar los *errores* o excepciones a la norma. Si escogemos la moda como representación del conjunto de datos, sabemos que estamos acertando totalmente en tantos casos como sea la frecuencia de la moda, pero erramos en el resto. En el ejemplo sobre el color del cabello, al escoger el valor "Rubio", acertamos con exactitud en 20 ocasiones, y cometemos error de representación en los 30 casos restantes. Pero consuela saber que la moda es el valor que menos error genera, puesto que es la opción que menos datos deja fuera de la representación exacta. Si en lugar de "Rubio" escogiéramos cualquier otra categoría, la frecuencia de errores aumentaría.

De aquí surge una necesidad imperiosa: jamás hemos de utilizar sólo un índice o una medida de representación sin que vaya acompañada de una medida o un índice de *bondad de la representación*. En otros términos, necesitamos un recurso que acompañe a la moda y que exprese cuán buena es para representar al conjunto completo de datos. Una opción muy fácil e intuitiva es utilizar el porcentaje. Y eso haremos. Nuestra norma va a ser, por tanto, que ante variables nominales, la medida de representación va a ser Mo, mientras que la medida de bondad de la representación (que simbolizaremos con BMo) va a ser su porcentaje. Luego:

$$Mo = X_i, \text{ tal que } f_i \text{ es el máximo}$$

$$BMo = \%_i$$

En nuestro ejemplo, Mo=Rubio, BMo=40%.

Matizaciones

Cuando una variable nominal cuenta con muy pocas categorías, no tiene mucho sentido utilizar representaciones numéricas. Basta con escoger la tabla o la gráfica. Los casos tópicos corresponden a las variables dicotómicas, es decir, que cuentan con dos únicos valores. Por ejemplo: respuesta “Sí” o “No”, sexo “mujer” u “hombre”, hábitat “rural” o “urbano”, etc.

En ocasiones, por el contrario, se manejan muchas categorías en una misma variable nominal. En tales casos, la tabla de frecuencias tal vez no sea buena opción. Imagina, por ejemplo, que preguntamos a todos los estudiantes de psicología de una provincia, cuál es su lugar de nacimiento. A quienes nacieron en España le preguntamos por su provincia. En el caso de los extranjeros, para no generar excesiva dispersión, le preguntamos sencillamente su país, y lo consideramos como una provincia. Imagina que contamos con 60 categorías diferentes, tras haber entrevistado a 250 personas. Una tabla con 60 filas es inmanejable: demasiada información. De mismo modo, un ciclograma es una mala decisión, puesto que la gran cantidad de sectores imposibilita una buena interpretación del resultado. Lo mismo podemos decir respecto a una gráfica con 60 barras. Pensemos incluso en la representación numérica. La moda tiene sentido, puesto que selecciona información relevante. Pero cabe esperar que BMo tenga un valor demasiado pequeño como para que Mo nos genere suficiente credibilidad.

Cuando contamos con muchas categorías, lo mejor es *fundir*, crear categorías nuevas que surgen de unir las frecuencias de otras. En el ejemplo del lugar de nacimiento, supongamos que estamos preguntando en la facultad de psicología de la Universidad de Sevilla. Cabe esperar que Sevilla sea la provincia más frecuente, seguida por otras de Andalucía y que, conforme nos alejamos, las frecuencias disminuyan. Una solución para la abundancia de categorías es considerar algo así como: (1) Sevilla, (2) Otras provincias andaluzas, (3) Otros lugares de España, (4) Otros lugares de Europa, (5) Resto del mundo. De este modo, hemos reducido las 60 categorías iniciales a 5. Cuando llevamos a cabo una acción de estas características es recomendable hacerlo constar en el informe: la variable original se expresaba de tal modo, pero recodificamos, generando la siguiente distribución de valores, mediante estos criterios.

Variable ordinal

Escala

Las variables ordinales son muy frecuentes en psicología. Bajo esta etiqueta se pueden identificar, además, grados diferentes de conseguir “finura” o “precisión” en las mediciones. Ambas circunstancias hacen que, si bien estrictamente hablando una variable ordinal es cualitativa, en muchas ocasiones puede considerarse “casi” cuantitativa o cuasicuantitativa. Esta distinción es trascendente. Las herramientas de análisis de datos disponibles para variables cuantitativas son mucho más abundantes, versátiles, extensas en aplicaciones y potentes que las herramientas disponibles para otros tipos de variables. Luego, cuando contamos con una ordinal con un elevado nivel de medida, en psicología solemos abordarla como si fuera cuantitativa, lo que reporta buenos resultados. Vamos a abordar primero su definición general, para distinguir después las *cuasicuantitativas*.

Una variable ordinal es aquella que apunta a estados tales que pueden ser ordenados. Pensemos, por ejemplo, en la pregunta “¿Cuánto frío sientes esta mañana?”, cuyas respuestas posibles son: (1) ninguno, (2) casi nada, (3) cierto fresco, (4) mucho, (5)

me muero de frío. Está claro que “mucho” es más que “cierto fresco” y, por tanto, procede considerar que el valor 4 es mayor que el valor 3 y no sólo valores distintos como quedaba resuelto el tema para las variables nominales.

Veremos que en el caso de las variables cuantitativas, no sólo podemos ordenar sino suponer la existencia de una cuantía o unidad de medida básica. No ocurre así en las ordinales. Entre “casi nada” y “cierto fresco” no hay algo parecido a 3 ó 4 *sensifresquinos* o cualquier otro invento. Por ello no podemos suponer que la distancia en sensación de frío que existe entre los valores 3 y 4 (cierto fresco y mucho, respectivamente) sea la misma que existe entre 2 y 3 (casi nada y cierto fresco, respectivamente), aunque estemos tentados a ello al tratarse de la misma diferencia aritmética (4-3 = 3-2 = 1). Así pues, en una variable ordinal sólo podemos decir que 4 es más que 3, pero no cuánto más.

En algunos procesos de medición en psicología, podemos llegar más lejos. Pensemos, por ejemplo, en una medida de ansiedad donde se ha respondido a un conjunto de veinte items, cada uno de ellos medidos en un continuo de 7 puntos donde cada persona ha respondido pensando en cuán cerca se encuentra su situación entre los dos extremos (1 y 7) de posibles respuestas. Se ha llevado a cabo una investigación concienzuda para garantizar que las personas entienden bien y del mismo modo cada enunciado. Se han realizado estudios para interpretar correctamente las puntuaciones que se derivan del test. En estas condiciones, los resultados de aplicación del cuestionario estandarizado de ansiedad son muy *finos* o *precisos.* El conjunto de datos resultante puede ser considerado como si fuera una variable cuantitativa, es decir, una ordinal *fina* o variable *cuasicuantitativa*.

El problema básico en la escala ordinal es suponer que la persona que responde está o no utilizando una regla o escala de unidad constante a la hora de facilitar el dato. Observa el continuo siguiente:

```
Nunca – Algunas veces – Bastantes veces – Muchas veces – Siempre
```

Son las respuestas que ofrecemos a un conjunto de personas para responder a la pregunta “¿Con qué frecuencia piensa usted que tendrá problemas con su vehículo cuando circula con él?”. Cada persona puede interpretar cada una de las categorías de respuesta de un modo diferente. Así, por ejemplo, hay algunas personas para quienes “bastante” es más que “mucho”, por lo que incluso se viola el principio de constancia en el orden. Estamos entonces en una caso de variable ordinal sin apellidos (ordinal, ordinal burda u ordinal no cuantitativa).

La misma pregunta podría ser utilizada de otro modo. Podemos definir sólo los dos extremos de respuesta, como se hace por ejemplo en la siguiente instrucción: *Responda con un número entre 0 y 10, donde 0 representa “nunca” y 10 “siempre”*. En este caso, la persona utiliza una especie de regla interna, de tal forma que podemos suponer (con cierta asunción de error), que el valor 7 tiene un significado similar en términos de frecuencia en todas las personas, y que 6 viene a ser el doble de veces que 3 en todas las personas. En términos generales, un argumento para suponer existencia de *cuasicuantitativa* es utilizar formatos de respuesta donde no se recurre a etiquetas para las categorías sino que sólo se explicitan los extremos, forzando a que la persona que responde utilice su regla interna. Esto es más contundente conforme mayor sea el número de alternativas de respuesta que se ofrezcan.

Tabulación

La tabla correspondiente a una variable ordinal tiene el mismo objetivo, sentido y dinámica que en el caso de una variable nominal, salvo en un aspecto: ya que hay orden, aprovechémoslo. Ocurre que en muchas ocasiones interesa responder a preguntas que se refieren a un intervalo de valores. Por ejemplo, ante los resultados en un examen, podemos preguntar por el número de personas que han aprobado. Si la evaluación se ha localizado en el intervalo de 0 a 10 y el aprobado se consigue a partir del 5, inclusive, la pregunta es tanto como interrogar sobre cuántas personas han obtenido el valor 5 o mayor. En el ejemplo de la frecuencia que he utilizado en el apartado anterior podríamos preguntar por cuántas personas piensan que van a tener problemas conduciendo, como mucho, bastantes veces, lo que implica acumular las frecuencias de las respuestas “nunca”, “algunas veces” y “bastantes veces”. Vamos a observar primero una versión de tabla de frecuencias tal como la hemos conocido para la variable “color de cabello” pero referida a “frecuencia de pensamientos negativos conduciendo”:

X_i	f_i	$\%_i$
1	20	25
2	32	40
3	16	20
4	8	10
5	4	5
Σ	80	100

Para saber cuántas personas piensan que van a tener problemas conduciendo en bastantes ocasiones, como mucho, hemos de *acumular* las frecuencias 20, 32 y 16 (68 personas) o los porcentajes 25, 40 y 20 (85%). Dado que podemos generar varios interrogantes similares, es una buena decisión incluir esta información en la tabla, en forma de *frecuencias absolutas acumuladas* (Fi) o *porcentajes acumulados* (%ai):

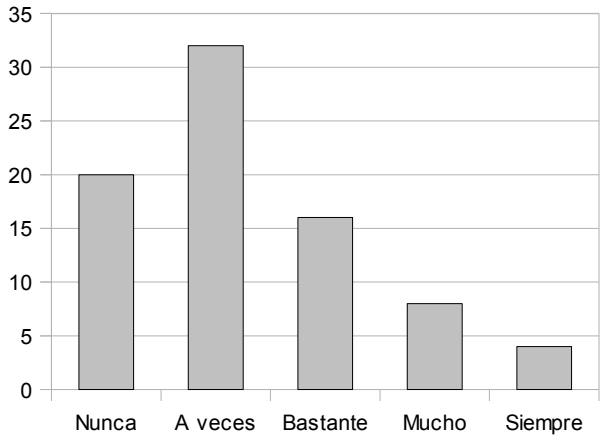
X_i	f_i	$\%_i$	F_i	$\%a_i$
1	20	25	20	25
2	32	40	52	65
3	16	20	68	85
4	8	10	76	95
5	4	5	80	100
Σ	80	100		

Gracias a esta información sabemos que un 65% de las personas entrevistadas piensa que nunca o sólo algunas veces llega a considerar que le pase algo conduciendo. También podríamos concluir que un 35% (100%-65%) tiene pensamientos negativos conduciendo al menos bastantes veces. Esta información es relevante en variables ordinales. Recordemos que no es viable trasladar la misma mecánica a las nominales, puesto que en estas el orden no tiene sentido.

En algunas ocasiones, una variable ordinal o cuasicuantitativa (más en este segundo caso) muestra muchos valores diferentes, demasiados como para hacer aconsejable construir una tabla de frecuencias tal y como la estamos conociendo. Existen varias alternativas. Una, inmediata, es abandonar la idea de la tabulación y recurrir a una representación gráfica. Otra es utilizar intervalos de valores. Así, si el recorrido de la variable va de 0 a 50, podemos utilizar intervalos de 10 en lugar de valores: la primera fila de la tabla expresa la frecuencia de datos con valores comprendidos entre 0 y 5, la siguiente entre valores superiores a 5 hasta 10, etc. Existen también otros recursos, como el diagrama de tronco y hojas, pero exceden los objetivos de nuestro temario de conocimientos. El mejor consejo es que si tienes demasiados valores como para que una tabla de frecuencias sea manejable, recurre a una buena representación gráfica. Es lo que abordamos ahora mismo.

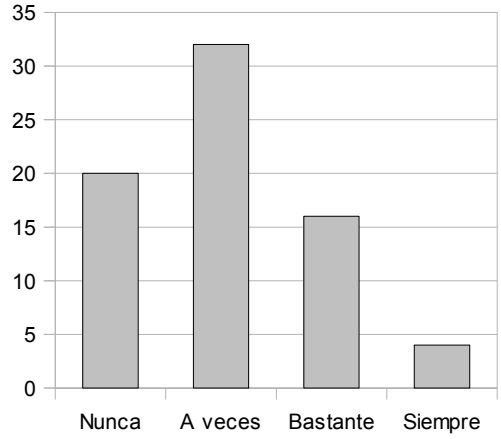
Representación gráfica

La mejor opción para una variable ordinal es el diagrama de barras. Las ventajas que el diagrama de barras tiene para las variables nominales son extensivas a las ordinales, si bien no ocurre lo mismo con el inconveniente: en este caso, que el diagrama de barras genere una sensación de orden es algo deseable y positivo. Si representamos los datos de la tabla anterior, lo que obtenemos es:

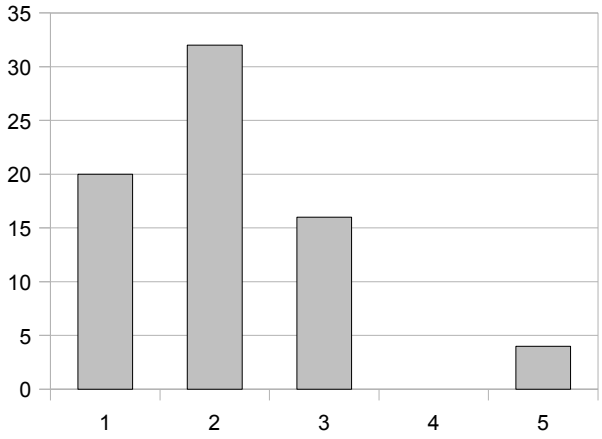


Se observa con rapidez que, salvo en el arranque “Nunca”, conforme aumenta la frecuencia de veces que alguien piensa que tendrá problemas conduciendo, disminuye el número de personas que dicen pensar así.

Recordemos que en una variable ordinal la cuantía concreta de cada valor es algo insustancial, lo importante es el lugar que ocupa esa cuantía. La lista anterior de las categorías podría ser codificada como he expresado más atrás: 1, 2, 3, 4 y 5. O bien podría ser objeto de otras cinco cantidades, como 1, 12, 27, 32 y 418. En sentido estricto da lo mismo, puesto que lo importante no es la cuantía sino el orden. Nos estamos situando en las variables ordinales puras, burdas o simples, no en las cuasicuantitativas. Extendiendo esta reflexión, ocurre que si hay una categoría que no se observa, sencillamente su valor numérico desaparece sin mayores consecuencias. Imaginemos que nadie escogió la categoría “Mucho”. En tal caso, la representación anterior sería algo así como:



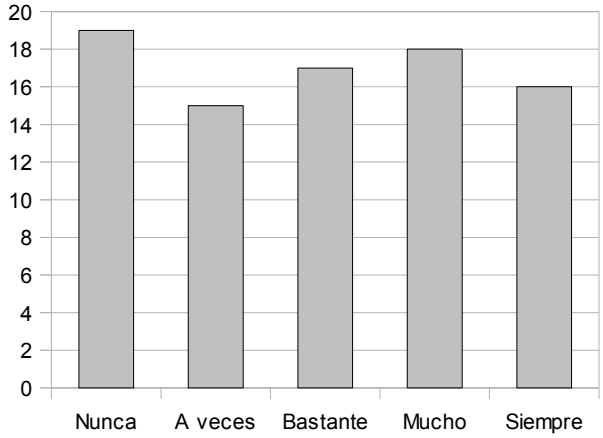
Cuando la variable es casi cuantitativa, los huecos son importantes. Imaginemos una pregunta similar a la que estamos trabajando, pero donde hemos indicado sólo los dos extremos: 1 representa ausencia total de sensación de frío, mientras que 5 representa un frío muy pronunciado. No etiquetamos los puntos intermedios. La representación gráfica que deberíamos observar en este segundo caso sería:



Cuando estamos manejando algo parecido a una regla donde se sitúan los valores de la variable, los huecos son fundamentales para la interpretación. Vamos a verlo más despacio en el contexto del tipo de variable donde esta afirmación es más relevante: las cuantitativas.

Representación numérica

Con las variables ordinales podemos utilizar también Mo como medida de representación del conjunto de datos. No obstante, dado que estamos manejando orden, es más que recomendable acudir a esta información y contemplarla en el índice de representación. Observa, por ejemplo, la siguiente representación de un conjunto de datos. En ella, Mo=”Nunca” y BMo=22,35%. Aunque sea el valor que menos frecuencia de error provoca, coincidiremos en que representa muy mal, puesto que se encuentra en un extremo de una distribución ordenada de valores con frecuencias nada despreciables.



Una buena estrategia para aprovechar el orden de los datos es, sencillamente, ordenarlos. Pensemos en una situación concreta. Hemos preguntado a 40 personas que evalúen de 0 a 10 la conducta de un personaje público concreto, donde 0 expresa claro desagrado y 10 una clara simpatía. Los resultados pueden ser los siguientes:

```
0, 2, 6, 3, 8, 1, 5, 7, 8, 1,
8, 4, 6, 8, 3, 4, 4, 4, 2, 4,
6, 9, 1, 9, 0, 5, 2, 8, 3, 5,
 2, 7, 6, 3, 9, 10, 8, 7, 6, 7
```

La primera tarea que vamos a realizar es ordenar los datos. El resultado es:

```
 0 0 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 4 5 5
5 6 6 6 6 6 7 7 7 7 8 8 8 8 8 8 9 9 9 10
```

Con sólo el orden, el conjunto de 40 datos permite una impresión más clara para comenzar a generar conclusiones sobre cómo se distribuye la variable. Estamos buscando una medida de representación del conjunto. Una buena idea es escoger el centro del listado ordenado. Es una medida intuitiva que permite resolver el problema que se provoca al escoger un valor que se acerca a un extremo y, por tanto, representa mal al otro. Para no decantarse por ninguno de ambos polos, la estratégica salomónica y aristotélica es escoger ese punto medio. Lo denominamos *mediana* (Md).

Para calcular Md se atraviesan tres etapas. La primera consiste en ordenar los datos desde el que posee el valor mínimo hasta el que muestra el máximo. Eso ya lo hemos hecho. El paso 2 implica localizar el centro del listado ordenado. Y el tercero, leer el valor de esa posición central. Ese valor es la mediana.

Para cumplir con el paso 2 partimos de que el conjunto total de datos es n=40. La posición central debe dejar 20 datos a un lado y 20 al otro. En otros términos, debe dividir a la distribución en dos porciones con igual cantidad de datos, que ocupan las posiciones de la 1 a la 20 (una mitad) y de la 21 a la 40 (la otra mitad). La posición central se encuentra, pues, entre las posiciones 20 y 21. Debe ser, por tanto, la posición 20,5. Aunque no hay una posición decimal en sentido estricto, sí que es la que divide equitativamente el conjunto en dos partes con igual frecuencia. Es algo así como plantearse que buscamos el centro entre las posiciones 1 y 40, lo que implica hacer la operación (40+1)/2 = 20,5. En términos generales, la posición central de un conjunto de n datos será (n+1)/2.

Ya sabemos que Md es $X_{20,5}$. Ahora hemos de leer el valor de $X_{20,5}$. Al observar el listado de datos e ir contando las posiciones, encontramos que el dato anterior y el posterior a la posición 20,5 tienen el mismo valor (X_{20}=5 y X_{21}=5). Luego:

$$\text{Md} = X_{\frac{n+1}{2}} = X_{\frac{40+1}{2}} = X_{20.5} = 5$$

Lo que acabamos de llevar a cabo es el cálculo de Md en una de las tres situaciones en que podemos encontrarnos. Otra situación ocurre cuando los datos colindantes a la posición central tienen valores diferentes. En ese caso, tomamos la misma decisión que con la posición central: calculamos el valor que se encuentra a medio camino entre los dos colindantes. Imagina que X_{20} tuviera el valor 5 y que X_{21} fuera 7. En ese caso, la mediana tendría el valor intermedio, es decir, 6. La última de las tres situaciones es la más sencilla: el número de datos es impar. En ese caso, la posición central coincide con una real, por lo que basta con observar el valor del dato que se encuentra en esa posición. Si añadiéramos un dato más a nuestro conjunto, entonces n=41 y la posición central sería (41+1)/2=21, que deja por debajo de sí a 20 datos (de las posiciones 1 a 20), del mismo modo que por encima (de las posiciones 22 a 41).

La moda, dijimos, es el valor que minimiza la frecuencia de errores, es decir, el valor tal que hace mínima la cantidad de datos que no coinciden exactamente con la representación. Pues bien, la mediana minimiza la *suma de los errores*. En otras palabras, si escogemos un valor cualquiera como representación del conjunto de datos y calculamos la distancia de cada dato con respecto a ese valor, la suma final será la más pequeña posible cuando ese valor de representación se trate de la mediana. Es una propiedad muy interesante, puesto que buscamos una representación que haga los errores o excepciones a la norma lo menos relevantes que sea viable.

Para la moda contamos con una medida de cuán buena es: su frecuencia en términos de porcentaje. Esta medida indica también indirectamente el grado de error cometido: conforme mayor sea el porcentaje, mejor es la moda como representación del conjunto, es decir, menor es el error que se comete al utilizarla en lugar de todo el conjunto de datos.

Para la mediana haremos lo mismo: una medida de su bondad de representación que, conforme muestre una cuantía mayor, indicará menor error cometido al utilizar su valor en lugar de todo el conjunto de datos. Para idear esta medida partimos de lo que acabamos de afirmar: las distancias a la mediana constituyen el conjunto de distancias de menor cuantía posible. Operamos, entonces, con las distancias. Pero ¿qué hacemos con ellas? Pues, coherentemente, podemos escoger, como medida de representación del conjunto de las distancias, su mediana. En otros términos, la bondad de la mediana, el valor que expresa en qué medida la mediana es una buena representación del conjunto de datos, será la mediana de las distancias a la mediana. Suele hacerse referencia a ello con MAD.

El cálculo de MAD es fácil de explicar: una vez se tiene la mediana, se calculan todas las distancias y se utiliza este conjunto nuevo para volver a calcular una mediana. Vamos a hacerlo.

Reproduzcamos de nuevo el conjunto anterior de datos:

0 0 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 4 5 5
5 6 6 6 6 6 7 7 7 7 8 8 8 8 8 8 9 9 9 10

Como Md=5, vamos a calcular ahora la distancia de cada valor respecto a la mediana, obteniendo:

```
5 5 4 4 4 3 3 3 3 2 2 2 2 1 1 1 1 1 0 0
0 1 1 1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 5
```

Ahora procedemos a calcular la mediana de este nuevo conjunto de datos, lo que implica ordenarlos:

```
0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 5 5 5
```

Ya sabemos que la posición central de estas 40 distancias es la 20,5, de nuevo entre dos datos con el mismo valor ($X_{20} = X_{21} = 2$), por lo que MAD = 2.

Tenemos entonces que el desagrado (0) o simpatía (10) que sienten las personas respecto a un personaje público tiene como valor más representativo Md=5 y como bondad de la representación, MAD=2. Conforme mayor sea el valor de MAD, peor será Md para representar al conjunto de datos. No hay una norma para considerar si el valor de MAD es tolerable o excesivo, pero podemos establecer (como haremos con el caso de las variables cuantitativas) que si MAD no supera el 50% de Md, Md es una buena representación numérica, mientras que si MAD=Md, la dispersión es muy elevada y Md no es una buena medida para representar a todo el conjunto de datos. Entre 50% y 100%, vamos a dejarle algún trabajo al sentido común.

En la literatura estadística, lo habitual es referirse a Md como una medida de *tendencia central* y a MAD como una medida de *dispersión*. Se habla de tendencia central bajo la perspectiva de que los datos tienden a ese centro. Dispersión es un término muy acertado para la bondad de la medida de representación, puesto que a más dispersión, peor será el recurso de utilizar un solo valor para sustituir a todos.

Matizaciones

MAD requiere el cálculo de algunas operaciones aritméticas que se llevan a cabo con mayor comodidad si la variable es ordinal cuasicuantitativa que si es ordinal burda. Es cierto que opera con el orden de las distancias, aunque para establecer ese orden debe considerar las cuantías. Es pues una medida que utiliza tímidamente cuantías, lo que es acorde con una cuantía también tímida por parte de las variables ordinales. No obstante, cuando esta posea pocos valores y su medida sea claramente muy burda, tampoco sería recomendable utilizar MAD.

No es fácil que los programas de ordenador utilicen MAD. Los hay que sí. Los hay que no. Es muy frecuente recurrir a otros índices de bondad de la representación o de dispersión en lugar de MAD aunque sean más defectuosos. Vamos a mencionar tres, los más habituales.

La medida más inmediata de dispersión es calcular la distancia que hay entre el valor más pequeño (mínimo) y el más grande (máximo). Aunque nos sorprenda, a esta medida de amplitud calculada con todos los datos se la denomina *Amplitud Total* (At). Es un recurso especialmente malo. Es medianamente habitual que las distribuciones de datos muestren valores raros en sus extremos. Si medimos, por ejemplo, altura, es posible que la gran mayoría de las personas del conjunto se encuentren entre alturas comprendidas entre 160 y 200 centímetros. Pero tal vez tengamos a alguien con 145 cm y a otra persona con 225 cm. Al considerar los extremos reales, el valor de At resulta inflado. Sería mejor despreciar un porcentaje de los extremos y calcular la distancia entre los valores mínimo y máximo de un intervalo central que ignore ese porcentaje lateral. Por ejemplo, podemos calcular la amplitud del 50% central, que desperdicia un 25% superior de los datos y un 25% inferior.

Ya veremos en otro momento que estos valores reciben el nombre de *cuartiles*. Es así porque son puntos de corte que dividen al conjunto ordenado de datos en cuatro partes con igual frecuencia o cantidad de datos. El cuartil primero (Q_1) deja por debajo de sí al 25% de los datos y por encima al 75% restante. El segundo (Q_2) deja a cada lado un 50% de los datos. A este segundo cuartil lo hemos bautizado ya como mediana (Md). El tercero y último (Q_3), deja por encima al último cuarto de los datos (25%) y a los tres cuartos restantes (75%) por debajo. Una medida de dispersión que ignore los cuartos extremos consistirá en calcular la amplitud que existe entre los extremos del 50% central, es decir, entre los valores de los cuartiles primero y tercero. Esta medida suele denominarse, también asombrosamente, *amplitud intercuartil*, $IQR=Q_3-Q_1$.

En ocasiones podréis encontrar otra medida parecida, que recurre a la mitad de esta distancia ($Q = [Q_3+Q]/2$) y que, con el mismo asombro al que estamos acostumbrados, se denomina *amplitud semi-intercuartil*.

Variable cuantitativa

Escala

Sin lugar a dudas, la cuantitativa es la reina de las variables. No es para menos. Contiene la máxima información que es capaz de explotar el análisis de los datos. Los números se han inventado para ellas. Hemos asumido, durante toda nuestra trayectoria de aprendizaje matemático, que el 8 es el doble que el 4 o que entre 3 y 7 existe la misma distancia que entre 11 y 15. Estas afirmaciones son ciertas sólo si existe una unidad de medida constante. Pensemos, por ejemplo, en el metro como unidad de medida de la característica “longitud”. 8 metros es el doble de longitud que 4 metros porque si tomamos un objeto que mida exactamente 1 metro, estará contenido en la primera distancia 8 veces, mientras que lo estará 4 veces en la segunda. Alguien mide 158 centímetros de altura porque ésta coincide con una pila de 158 piezas de 1 centímetro, una encima de otra.

La existencia de unidades de medida justifica el conjunto de las operaciones aritméticas. El sueño, pues, de alguien que opera con herramientas de análisis de datos es contar con variables cuantitativas. El sueño, no obstante, es muchas veces un sueño, sin sustento real.

En el campo de la psicología, lo estrictamente cuantitativo es una excepción. La norma está formada principalmente por variables nominales y ordinales. Algunas de estas, no obstante, surgen de procedimientos tan depurados y precisos que las consideramos cuasicuantitativas. Ya hemos hablado de ello.

Tabulación

Cuanto hemos aprendido sobre la tabulación de las variables ordinales es aplicable para las cuantitativas. No obstante, hay que considerar dos aspectos importantes. El primero es que podemos estar operando, en algunas ocasiones, con variables consideradas *continuas.* En sentido estricto, nunca manejamos variables cuantitativas continuas puesto que los instrumentos de medida son limitados y siempre recurren a una unidad mínima finita. Lo continuo implicaría que entre cualesquiera dos valores, siempre cabría un valor intermedio. Esto no ocurre en la práctica. Podemos estar manejando metros como unidad de referencia y llegar hasta los milímetros gracias a la precisión de la regla que utilizamos, pero no más allá. En tal caso, las medidas tendrán tres dígitos

decimales (decímetros, centímetros y milímetros). Aún así, cuando la variable es considerada continua por su riqueza de información, es necesario operar con intervalos y no con valores directos, como hemos visto ya en la tabulación de variables ordinales.

El otro aspecto relevante es que resulta muy frecuente que el número de valores que se manejan de las variables cuantitativas sea muy elevado. Una solución es seguir operando con intervalos, como en el caso anterior. Pero lo más recomendable es evitar la tabla de frecuencias y acudir a una buena representación gráfica. Recordemos que la función de la tabla es mostrar la información organizada de tal manera que sea fácil procesarla directamente mediante observación. Cuando la tabla muestra demasiada información, cuando alberga demasiados números, demasiadas cantidades, entonces pierde buena parte de su utilidad y hay que pasar a otro recurso.

Dejar a un lado la tabla para acudir a una gráfica no es la norma a seguir. En muchos casos, la variable cuantitativa se articula con un número cómodo de valores. Ocurre, por ejemplo, con el número de hijos de una familia, el número de horas diarias frente al televisor, el número de libros leídos en un mes, etc. La norma es “tabúlese salvo que deje de ser útil”. Aquí no hemos cambiado nada.

Lo fundamental en una variable cuantitativa es la característica mencionada de la unidad de medida. A la hora de interpretar esa variable es imprescindible utilizar una tabla o una gráfica que respete una regla exacta, una escala con una unidad continua. Imagina, por ejemplo, la siguiente tabla referente a la variable “Número de llamadas telefónicas realizadas por una muestra de 10 personas durante la última semana”

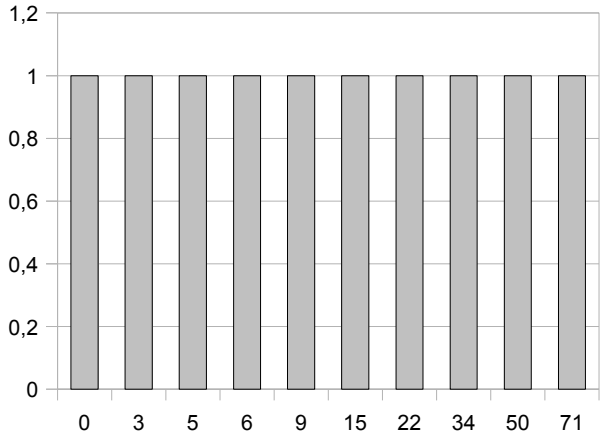
X_i	f_i	$\%_i$	F_i	$\%a_i$
0	1	10	1	10
3	1	10	2	20
5	1	10	3	30
6	1	10	4	40
9	1	10	5	50
15	1	10	6	60
22	1	10	7	70
34	1	10	8	80
50	1	10	9	90
71	1	10	10	100
Σ	10	100		

En primer lugar, es importante asumir que una muestra de n=10 no es lo habitual. Al aumentar n seguro que encontraríamos más variedad en los valores. En segundo lugar, aunque la información es interesante, dado que existen casi tantos datos como valores, la tabla no añade mucha utilidad frente, por ejemplo, a una inspección visual del conjunto completo de datos sin tabular. En tercer lugar, podemos extraer conclusiones erróneas demasiado fácilmente. Cualquier interpretación es difícil, puesto que existen importantes y variables *huecos* entre los valores. Observa, por ejemplo, que si bien hay quien realiza 5 llamadas y también hay quien hace 6, sin embargo de 50 a 71 hay un salto de 21 llamadas. Es complicado intentar hacerse una idea de las frecuencias de llamada observando una tabla de frecuencias habitual. O bien construimos intervalos o, mucho

mejor, desestimamos la tabla y realizamos una representación gráfica, conclusión tanto más contundente cuanto mayor sea *n*.

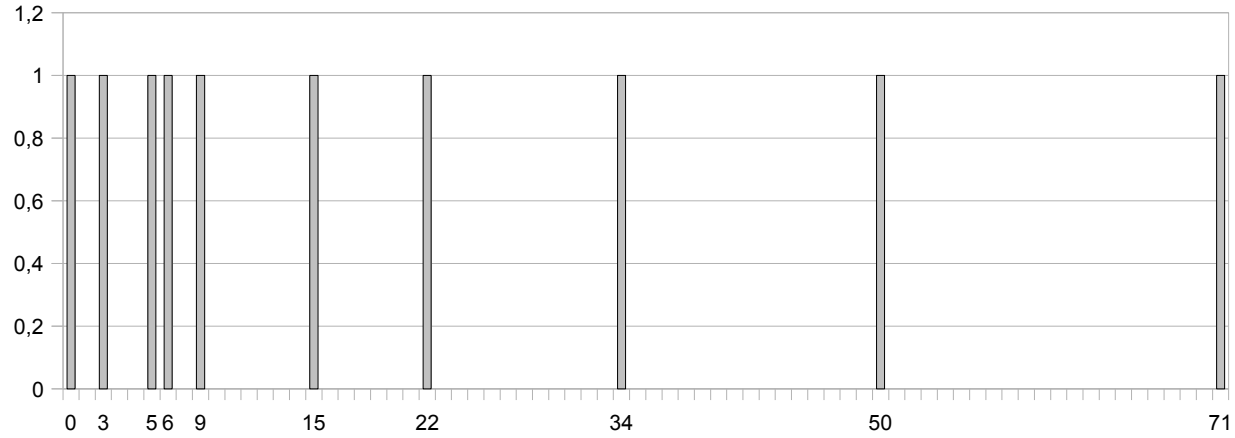
Representación gráfica

Para construir las mejores opciones en el caso de una variable cuantitativa vamos a partir del diagrama de barras. Lo hacemos a partir de la tabla de frecuencias anterior. Si se representa mediante el recurso comentado, tenemos este resultado:



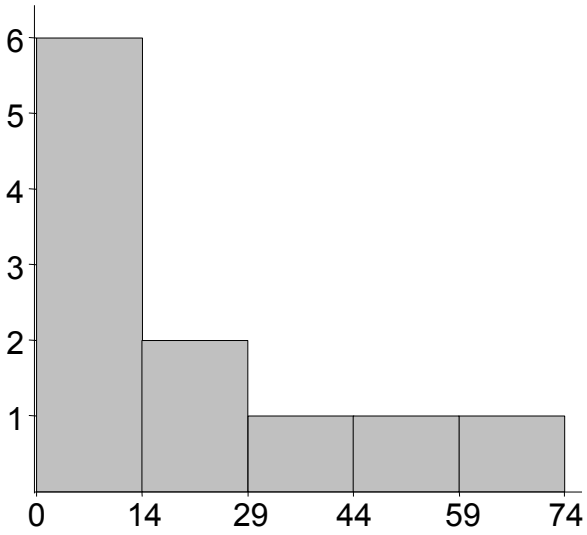
No sé qué pensarás tú. Para mí, esto es inútil, no sirve para nada porque no añade nada. Una situación habitual todavía es más contundente que esta. Imagina, por ejemplo, la representación mediante diagrama de barras del número de habitantes de los países del mundo. No hay dos con el mismo número, así que contaríamos con una ristra casi interminable de barras con la misma altura. Es peor que no hacer nada.

Tenemos que aplicar dos cambios importantes a esta representación. La primera es respetar que se trata de una variable cuantitativa, es decir, donde se respeta o debe respetarse una escala o regla de referencia. En otras palabras, los huecos son fundamentales para interpretar el resultado. Vamos a incluirlos en la representación, generando lo que sigue:



Los datos son exactamente los mismos, pero la interpretación que sugiere la gráfica no coincide con la que surgiría tras observar el diagrama de barras anterior. Gracias a que hemos respetado la regla de referencia o la escala cuantitativa, la

existencia de los huecos permite obtener una idea acerca de cómo se distribuye la variable en todo su recorrido. Una conclusión inmediata es que la frecuencia de llamadas tiende a agruparse ligeramente en los valores bajos y se extiende o dispersa hacia los valores altos. Esta misma conclusión podría ser más rápida si llevamos a cabo un segundo cambio. Lo que hacemos es considerar intervalos de valores. Es como comprimir el espacio horizontal e ir contando las unidades que nos encontramos por el camino. El recorrido de la variable va desde 0 hasta 71. Hemos de conseguir un número de intervalos con amplitud constante. Pongamos de 15 unidades. De este modo, conseguimos la siguiente representación.



En este caso, la interpretación resulta aún más sencilla. Tenemos, por ejemplo, 6 datos con valores comprendidos entre 0 y 14. Esta cantidad va decreciendo progresivamente mientras la frecuencia de llamadas telefónicas aumenta. Observemos en qué medida esta interpretación es mucho más sencilla en esta última herramienta gráfica, respecto al camino que iniciamos en la tabla de frecuencias.

Este recurso se denomina *histograma*.

Además del histograma contamos con diversas estrategias para representar gráficamente una variable cuantitativa. Tampoco voy a ser exhaustivo en este caso, pero antes de finalizar el apartado veamos una estrategia gráfica que nos acompañará en numerosas ocasiones, especialmente por su potencial para comparar grupos de datos cuantitativos entre sí: el diagrama de caja y patillas o, sencillamente, *diagrama de caja*.

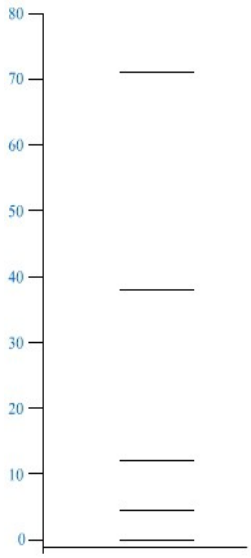
Para conocer con precisión cómo se construye y utiliza necesitamos información que todavía no manejamos con soltura. A pesar de ello, y ya que es este el momento en que hay que abordar los recursos gráficos, podemos tomar un primer contacto.

Para construir un diagrama de caja comenzamos identificando cuatro puntos en la distribución ordenada de datos:

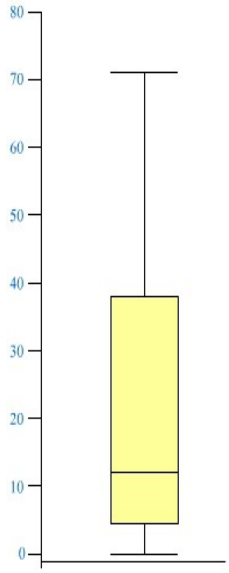
- Mínimo: el valor más pequeño del conjunto de datos.
- Cuartil 1: el valor de la posición que deja por debajo de sí al 25% de los datos.
- Mediana.
- Cuartil 3: el valor de la posición que deja por encima de sí al 25%.
- Máximo: el valor más grande.

En el ejemplo de las llamadas de teléfono, estos cinco valores son, respectivamente: $\min=X_1=0$, $Q_1=X_{2,75}=4{,}5$, $Md=X_{5,5}=12$, $Q_3=X_{8,25}=38$ y $\max=X_{10}=71$.

El siguiente paso consiste en situar estos valores en una representación gráfica. Partimos de un eje horizontal donde se representan las variables, mientras que en el vertical se encuentran los valores. Cada uno de estos cinco puntos calculados quedan simbolizados mediante una línea recta en la gráfica.



En el siguiente paso, unimos las líneas 2 y 4 en una caja. Las líneas 1 y 5 se unen a la caja con la patilla. Observa cómo queda. Ya puestos, he coloreado la caja.



Los cinco puntos generan cuatro zonas, cuatro cuartos de la distribución ordenada de datos. La patilla inferior representa el 25% de los datos. Desde el límite inferior de la caja a la línea situada en su interior (mediana), encontramos otro 25%. Lo mismo entre la mediana y el borde superior de la caja, para terminar con el último 25% representado por la patilla superior. De esta descripción se concluye con facilidad que las patillas representan el 50% más extremo, mientras que la caja implica un 50% de los datos más significativos del conjunto. La conclusión que obtuvimos en la observación del histograma es literalmente transportable a este recurso de caja y patillas.

Representación numérica

En la clasificación que utilizamos no hay variable que contenga más información que la cuantitativa, así que los índices o medidas que se utilicen con estas deberían considerar toda la información. Sabemos que la moda sólo tiene en cuenta la frecuencia, tomando el valor más frecuente y sin recoger información sobre el resto del conjunto de datos. La mediana va más allá y absorbe también la posición. La cuantía de los datos es considerada únicamente como elemento fundamental para establecer orden, acto seguido se pierde.

La media aritmética permite, a diferencia de las anteriores, tener en cuenta la cuantía de todos los datos en un mismo recurso. La estrategia para construir el valor más representativo del conjunto de datos consiste en:

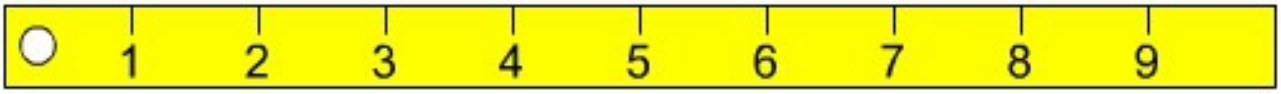
$$media\ aritmética = \frac{suma\ de\ datos}{número\ de\ datos} = \frac{\sum X_i}{n}$$

Observa que todos los datos dejan su huella, que resulta tanto mayor cuanto mayor sea también el valor del dato. Podemos pensar también en la media aritmética como en una repartición equitativa: es como si repartiéramos todas las cuantías de forma equitativa entre todos los datos.

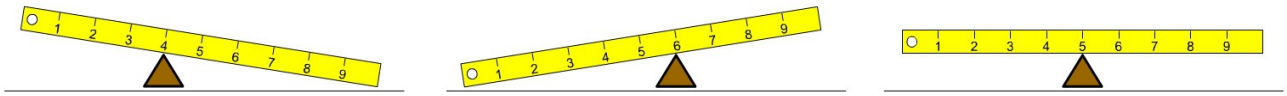
Como hemos visto, la moda es el valor que minimiza la frecuencia de errores y la mediana es la medida que minimiza la suma de errores. Pues bien, la media aritmética minimiza la suma de errores cuadráticos (errores al cuadrado). Esto parece una tontería (y, en cierto sentido, lo es), pero tiene su importancia en diversas aplicaciones, especialmente en el contexto donde triunfó inicialmente la aritmética basada en la media y sus derivados: la astronomía. Vamos a ver, además, que el instrumento utilizado como bondad de la media utiliza distancias cuadráticas y, por tanto, la media aritmética es el valor que minimiza el resultado de ese instrumento.

Detengámonos un momento en la relación que guarda la media aritmética con la representación gráfica de la variable. Para ello, vamos a construir una gráfica “ladrillo a ladrillo”.

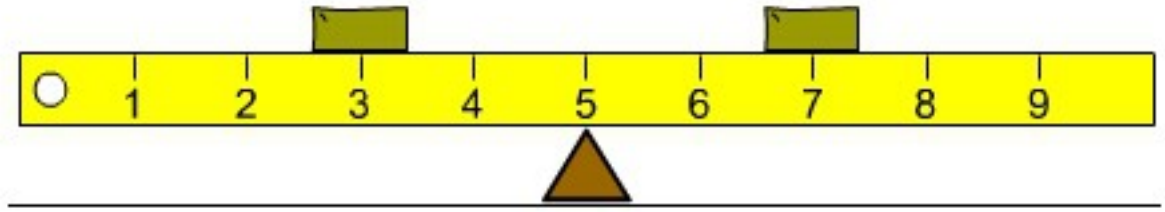
Lo primero que hacemos es disponer de una regla. Imagina que se trata de una regla física real, suficientemente rígida y fuerte como para aguantar sobre ella mucho peso, ya que vamos a ir colocando encima ladrillos:



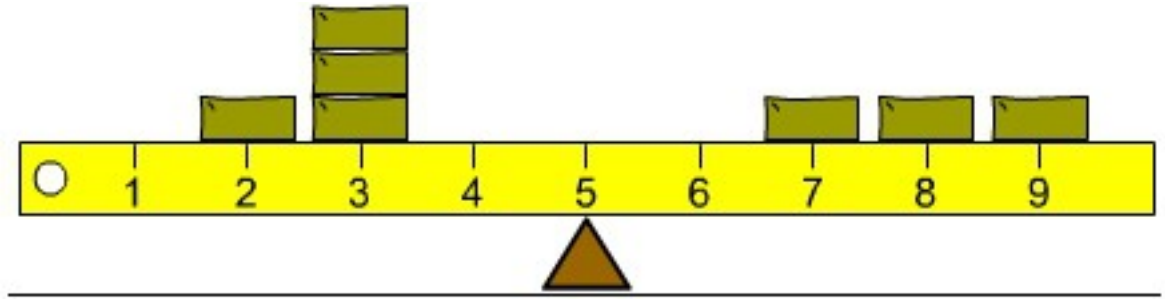
La regla mide exactamente 10 unidades (no importa si son centímetros, metros o lo que quieras imaginar, mientras sean unidades de longitud constante). Vamos a depositarla sobre suelo. Ya está. Ahora vamos a colocarla sobre un punto de apoyo, por ejemplo, un triángulo con la punta hacia arriba. ¿Dónde hay que colocar el triángulo para que la regla se encuentre en equilibrio? Lo que sigue es la representación gráfica de tres situaciones, donde se observa que la única posibilidad para conseguir que la regla se mantenga paralela sin tocar el suelo es situando el punto de apoyo exactamente en el centro: el valor 5.



Pongamos ahora algunos ladrillos. Si ponemos un ladrillo en la posición 3 (dos unidades por debajo del punto de apoyo), tendremos que colocar otro en la posición 7 (dos unidades por encima) para que la regla siga estando en equilibrio:



Sigamos con el mismo juego. Ahora vamos a añadir unos cuantos ladrillos más: uno en el 2 y dos en el 3. Tras estos tres ladrillos es obvio que la regla se inclinaría hacia la izquierda, así que pensemos en una solución. Cada ladrillo puesto en el 3 implica dos unidades de distancia respecto al punto de apoyo situado en el 5, lo que implica cuatro unidades de peso hacia la izquierda. El ladrillo colocado sobre el 2 añade tres unidades más de peso en la izquierda. Por tanto, tenemos una descompensación de siete unidades. Hemos de añadir otras siete a la derecha para que el punto de apoyo en 5 siga cumpliendo su función. Dado que 5+7=12 y que la regla sólo tiene escala hasta el 9, necesitaremos al menos 2 ladrillos cuyas distancias sumen 7. Por ejemplo, uno en 9 (9-5=4) y otro en 8 (8-5=3). De esta forma, el panorama queda como sigue:



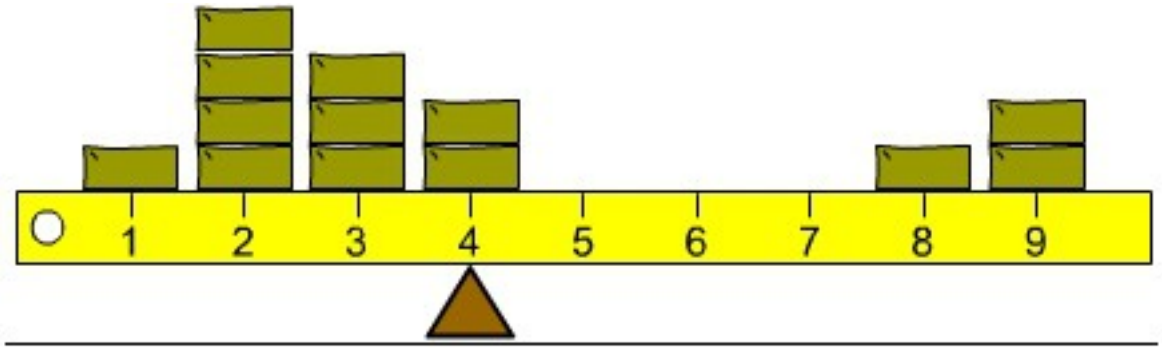
Lo que hemos hecho con los pesos, encontrando el punto de apoyo, tiene un significado central en el asunto de la representación en variables cuantitativas: *el valor numérico del punto de apoyo es la media aritmética*. En otros términos, del mismo modo que la mediana es el centro de frecuencias (deja a ambos lados la misma cantidad o frecuencia de datos, tengan estos más o menos peso), la media aritmética es el centro de pesos, masas o cuantías (deja el mismo peso a ambos lados, sea o no con la misma cantidad de datos). Vamos a insistir en ello con otro ejemplo.

Calculemos la media aritmética del siguiente conjunto de datos. Dado que la media es una repartición equitativa de las cuantías entre los datos, surgirá de dividir la suma de las cuantías entre el número de datos:

2, 3, 4, 3, 9, 1, 4, 8, 2, 2, 9, 2, 3

$$\overline{X} = \frac{\sum X_i}{n} = \frac{52}{13} = 4$$

La siguiente representación gráfica muestra el efecto: para que la regla se mantenga en equilibrio, el punto de apoyo debe estar situado exactamente en el valor de la media aritmética, 4. Si te implicas en comprobarlo, verás que las distancias a la media suman 14 puntos a ambos lados.



Este comportamiento como centro de pesos muestra un efecto o propiedad interesante, que motivará algunas decisiones acto seguido. Ocurre que si a ambos lados de la media aritmética se encuentra el mismo peso, masa o cuantía de distancias, entonces la suma de las distancias a la media debe dar siempre el mismo resultado: 0. Comprobemos que es exactamente lo que ocurre con los datos del ejemplo[1]:

$$(2, 3, 4, 3, 9, 1, 4, 8, 2, 2, 9, 2, 3) - 4 =$$
$$= -2, -1, 0, -1, 5, -3, 0, 4, -2, -2, 5, -2, -1 \rightarrow$$
$$\rightarrow \sum (X_i - 4) = 0$$

La media aritmética, en definitiva, será nuestra opción como medida de representación del conjunto de datos cuando estos sigan una escala cuantitativa. Del mismo modo que hemos hecho hasta el momento, necesitamos también para este índice una estrategia que exprese cuán bueno es para representar al conjunto de los datos. En el caso de la moda recurrimos a su porcentaje. Para la mediana, la medida de bondad preferible es el MAD. ¿Y para la media aritmética?

Dado que estamos operando con las cuantías de todos los datos para calcular la media aritmética, parece natural que operemos con las distancias cuantitativas de todos los datos a la media aritmética. El MAD es la mediana de las distancias a la mediana. Del mismo modo, lo que hacemos ahora es tomar como bondad de la media, o medida de dispersión de datos cuantitativos, la media de las distancias a la media:

$$Bondad\ \ media = \frac{\sum (X_i - \overline{X})}{n}$$

Problema: acabamos de razonar que la suma de las distancias a la media aritmética es siempre cero. Así que ese índice dará también el mismo resultado. Una forma aritmética de solucionar el problema es elevar las diferencias al cuadrado, puesto

1 Se puede demostrar que siempre ha de ocurrir lo mismo, pues:
$$\sum (X_i - \overline{X}) = \sum X_i - \sum \overline{X} = n\overline{X} - n\overline{X} = 0$$

que todo cuadrado es siempre positivo. Esta medida es coherente con la circunstancia de que la media aritmética minimiza los errores cuadráticos, es decir, con que la media aritmética es el valor de referencia que hace mínima la suma de distancias al cuadrado a un valor de referencia. Este índice o medida tiene el nombre de *varianza*, se simboliza con la letra S al cuadrado (S^2) y, como hemos dicho, se calcula mediante la expresión:

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

La varianza es un índice de dispersión muy utilizado para cálculos intermedios, pero cuenta con un importante inconveniente a la hora de interpretar su cuantía. Si estamos midiendo longitudes en metros, los valores originales se expresan en metros y la media aritmética también, por lo que la varianza se expresa en metros al cuadrado. Esto implica situaciones como la siguiente: "¿Qué puedes decirme sobre el número de hijos que tienen las familias de tu ciudad?" "Pues tienen 3,2 hijos por términos medio, con una varianza de 4,8 hijos al cuadrado". ¿Qué cosa es esa de un hijo al cuadrado?

Para solucionar el problema de la unidad de medida, utilizamos el recurso más inmediato: aplicar la raíz cuadrada. El resultado tiene un nombre y tres apellidos a escoger: desviación tipo, desviación típica o desviación estándar. Se simboliza con la letra S, recordando que es la inicial de la expresión inglesa *standard*. Por eso, porque la desviación tipo se simboliza con una S, su cuadrado (la varianza) se expresa con una S^2.

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}$$

En el ejemplo anterior, los valores de la varianza y de la desviación tipo son, respectivamente, 7,23 y 2,69.

Matizaciones

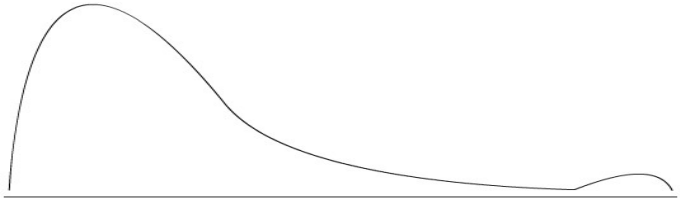
Uno de los puntos más relevantes de este documento es que una medida de representación del conjunto de datos no puede suministrarse o manejarse sin ir acompañada de otro recurso: una medida de dispersión o de cómo de buena es la representación. En términos generales hemos afirmado que:

entonces utilizamos como	*Cuando la variable es del tipo:*		
	Nominal	**Ordinal**	**Cuantitativa**
representación:	moda	mediana	media
bondad de rep.:	% de la moda	MAD	desviación tipo

Como ya hemos visto en el caso de las representaciones gráficas, las reglas generales son tan útiles como imperfectas y existen diversas situaciones en las que hay que poner en práctica excepciones. El objetivo general es siempre el mismo: el instrumento debe servir al objetivo y no convertirse en un objetivo en sí mismo o en un comportamiento irreflexivo.

Imagina que hemos obtenido cuál es el ingreso mensual de un conjunto de mil personas. Las variables que se refieren a la renta suelen ser muy dispersas, especialmente respecto a los valores altos (unos pocos valores con rentas muy altas). Al representar esta variable en un histograma podríamos encontrar algo parecido a la gráfica

siguiente. Observa que se muestra una clara asimetría positiva. Es una distribución asimétrica porque muestra un comportamiento extraño o disperso en uno de los lados de la distribución. Y es *positiva* porque el *lado raro* se encuentra en la zona de los valores más altos, es decir de las diferencias positivas respecto a la media aritmética.



La mayor parte de la población tiene una renta que se sitúa en la parte baja de la gráfica, es decir, con salarios, ingresos o riquezas de valores relativamente bajos. No obstante, hay una pequeña parte de la población que tiene rentas de valores muy altos. La media es muy sensible a esas cuantías. Los salarios muy altos incrementarán sensiblemente la suma final, por lo que pocos datos generarán un desplazamiento sensible de la media aritmética, llegando a ser una pésima medida de representación del conjunto. Observa los siguientes datos:

A: 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3

Representan el grado de ansiedad de un grupo de catorce personas. El cuestionario admite puntuaciones hasta de valor 100. Digamos, entonces, que se trata de un grupo de personas que inicialmente parecen muy poco ansiosas. No obstante, preguntamos ahora a dos personas que acaban de ingresar en el grupo y obtenemos el siguiente resultado:

B: 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 90, 95

Calculamos ahora las representaciones numéricas basadas en el orden y la cuantía para los dos conjuntos de datos:

Conjunto	Representación		Bondad	
	Medida	Valor	Medida	Valor
A	Mediana	2,00	MAD	1,000
	Media	2,00	Des. tipo	0,845
B	Mediana	2,00	MAD	1,000
	Media	13,31	Des. tipo	29,954

Observa lo que ha ocurrido. Al añadir dos datos especialmente raros a la derecha del conjunto previo (en la posición alta), la media ha aumentado sensiblemente de valor y la desviación tipo se ha disparado. Esta última nos indica que la media es una representación pésima. Resulta particularmente raro que el valor de la desviación tipo sea superior a la media aritmética. En este caso resulta ser más del doble. Se observa, además, que el valor 13,31 no representa bien a la mayoría de los datos (que se

encuentran entre 1 y 3) ni a las cuantías raras (90 y 95), es decir, a ningún dato del conjunto.

Moraleja: cuando el conjunto de datos tiene una marcada asimetría, es decir, una dispersión sensible a uno de los lados de la distribución, la media aritmética y por tanto también la desviación tipo son recursos desaconsejados. ¿Qué hacer en estos casos?

Tenemos varias posibilidades. Veamos tres.

Una estrategia es prescindir de los extremos. Es un recurso *automático* al que se acude cuando se teme que ocurran cosas como la ejemplificada pero no se desea emplear un tiempo específico en estudiar la variable en profundidad. Una decisión frecuente cuando se opta por esta estrategia es considerar el 95% central de los datos y prescindir del 2,5% de cada extremo. Suponemos entonces que hemos podido controlar *rarezas* diversas.

La estrategia anterior puede ser mejorada sensiblemente mediante otras posibilidades. La segunda consiste en escoger otro índice de representación y, por tanto, otra medida de dispersión. Una buena decisión es acudir a la mediana. Observa en el ejemplo que se trata de una medida muy *resistente*, es decir, que no se deja afectar por las cuantías raras. Esta opción es muy frecuente y recomendable. Otra salida en este mismo sentido es acudir a unas medidas denominadas *M-estimadores*. Los M-estimadores se calculan mediante ciclos repetitivos donde se observa qué ocurre con una representación inicial que va modificándose cuando se consideran determinadas perturbaciones en los datos. A base de cálculos y recálculos va afinándose el mejor valor de representación. Sólo puede llevarse a cabo con un ordenador y su comprensión requiere cierto esfuerzo, así que ni vamos a estudiarlo ni se usa en la práctica salvo alguna excepción. Nuestra opción será considerar la mediana y, por tanto, MAD.

La tercera estrategia es estudiar los datos en dos fases. En la primera se desestiman los valores raros y se llevan a cabo las conclusiones con el resto, con la generalidad. Una vez concluidos estos análisis, se recuperan los valores desestimados y se lleva a cabo un estudio específico con ellos. Este procedimiento es muy recomendable en psicología, donde nos interesa cómo funcionan determinados efectos en general (por ejemplo, una intervención en una comunidad o una terapia con un paciente) tanto como el comportamiento particular de algunos casos, ya que todas las situaciones a las que se enfrenta un profesional de la psicología son únicas e irrepetibles si bien con una base común.

En el caso de la renta o de los ingresos, por ejemplo, si la asimetría es fluída o sin cambios bruscos, lo más recomendable es acudir a la mediana. Si se identifica con claridad un grupo o varios, entonces debería realizarse un análisis por separado para cada uno de ellos.

El coeficiente de variación de Pearson

Ya hablaremos del señor Pearson en otra ocasión. Verás que las historias de la vida real superan la ficción, incluso de las telenovelas con más éxito de audiencia. Nos encontraremos con esta persona en varios temas. Hoy aparece aquí porque nos ha legado una estrategia muy intuitiva y sencilla de calcular para interpretar la cuantía de la dispersión en variables cuantitativas.

Ya he insistido en que toda medida de representación numérica o tendencia central ha de ir acompañada de una estrategia que permita establecer en qué medida ese número es una buena representación del conjunto de datos. Hemos denominado a esta estrategia *bondad de la representación* o *medida de dispersión*. Como sabemos, en el caso de una variable ordinal, la mejor representación numérica es la media aritmética y la

mejor medida de bondad, la dispersión tipo. Pues bien, ¿qué cuantía de la dispersión me está diciendo que ya es demasiado y que la media no es una buena decisión? No tenemos una respuesta cerrada, como podrías temer. Recuerda que en el caso de la mediana y su MAD, hemos considerado que si MAD supera el umbral del 50% de la cuantía de la mediana, debemos comenzar a pensar en que representar todos los datos por ese número tal vez no sea una buena decisión. Tomemos esa misma solución: si el valor de la desviación tipo supera la mita del valor de la media aritmética, entonces comenzamos a pensar que hay mucha dispersión. Si se ha superado el 100%, entramos en el "demasiado" y lo característico no es la media, sino la desviación. Es decir, si alguien nos pregunta por cuál es el valor más representativo, nuestra respuesta debería ser "lo característico es la dispersión de valores". Y ahí queda la cosa.

Pearson propuso precisamente esa estrategia: expresar la desviación tipo como porcentaje de la media. Lo llamamos Coeficiente de Variación de Pearson:

$$CV_{Pearson} \text{ o, sencillamente, } CV = 100\frac{S}{\bar{X}}$$

Si bien CV es una estrategia recomendable para interpretar la cuantía de la desviación tipo, hay una situación en la que no es recomendable sino necesaria: cuando se comparan las dispersiones de dos o más conjuntos de datos de variables cuantitativas. Dado que cada conjunto es peculiar, las medias aritméticas pueden variar mucho entre sí y, por tanto, los mismos valores de la desviación tipo no se interpretan del mismo modo.

Imagina a dos grupos de personas conversando en el parque. Nos interesa conocer de qué edades son sus componentes. En el grupo que conversa junto a la fuente, la media de edad es de 50 años, con una dispersión de valor 8. El grupo que interactúa en la zona de los columpios tiene una edad media de 8 años, con una dispersión de valor 4. En principio, 8 es más que 4, por lo que podríamos concluir que hay más dispersión en el grupo de la fuente que de los columpios. Pero pensemos esto un poco más despacio.

Una dispersión de 8 años cuando hablamos de edades que rondan los 50 es, en términos relativos, un CV de valor 16 (la desviación tipo es un 16% de la media aritmética). Una dispersión de 4 años en un grupo de niños cuyas edades rondan los 8 años es una variabilidad notable. Estamos hablando de un CV del 50%. Pensemos en términos de una dispersión arriba o abajo de la media. En el grupo de niños, hablamos de edades entre 8-4=4 y 8+4=12. En el grupo de mayores, 50-8=42 y 50+8=58. Entre un niño de 4 años y un chico de 12 hay una sinfín de diferencias. Nos asombraría que el niño de 4 tuviera más información que el de 12 o fuera capaz de hacer más flexiones o de recitar una poesía mejor o... Pero no es extraño que una persona de 58 sea capaz de correr más tiempo o encontrarse en mejor estado de salud que alguien con 42. Las diferencias de edad son menos importantes conforme aumenta la edad. Luego, cuando comparamos dos variabilidades, que provienen de dos conjuntos de datos diferentes, es importante acudir a una medida relativa, como la CV, antes que a una absoluta, como la S.

El coeficiente de dispersión acotada

El coeficiente de variación de Pearson se encuentra ampliamente extendido como medida para *relativizar* la dispersión de una variable. No obstante, a pesar de sus ventajas, cuenta con dos inconvenientes de peso. El primero es que la dispersión puede llegar a ser muy superior a la media aritmética, por lo que el límite superior de CV no es 100. Para interpretar bien un índice nos interesa que esté acotado. CV no lo está y, por

tanto, su interpretación también es difícil. Observa el siguiente conjunto de datos. Hemos calculado un CV ¡de valor 279! Esto hace casi bueno un CV = 100.

X	f
1	35
2	10
3	3
10	1
50	1
	50

Media: 2,48
Desviación tipo: 6,92
CV de Pearson: 279

Otro inconveniente de CV es que se comporta mal cuando se combinan puntuaciones negativas y positivas, puesto que la dispersión no se ve afectada por ello, pero la media puede llegar a ser 0. Es más, si ese es el valor de la media, CV no puede ser calculado, no tiene solución.

Una solución es el índice de dispersión acotado. Consiste en:

- Encontrar cuál es el valor máximo que puede tener la dispersión, en las condiciones del conjunto de datos (considerando su recorrido y su media)
- Expresar la dispersión real como un porcentaje de la máxima.

Se puede deducir que el valor máximo que puede tener la varianza de una variable es el resultado de multiplicar la distancia de los extremos a la media aritmética, es decir:

$$máx(S^2) = (\bar{X} - mín)(máx - \bar{X}) \rightarrow máx(S) = \sqrt{(\bar{X} - mín)(máx - \bar{X})}$$

En el ejemplo anterior, el valor máximo que podría tener CV (con media de valor 6,92, mín=1 y máx=50) es

$$máx(CV) = \frac{100}{\bar{X}}\sqrt{(\bar{X} - mín)(máx - \bar{X})} = \frac{100}{6{,}92}\sqrt{(6{,}92 - 1)(50 - 6{,}92)} = 231$$

La desviación acotada (Sª) expresa la dispersión tipo real como un porcentaje de la desviación tipo máxima, haciendo:

$$S^a = \frac{100\ S}{\sqrt{(\bar{X} - mín)(máx - \bar{X})}}$$

En el ejemplo con CV = 279, Sª = 83. Observa el siguiente caso, en el que CV no puede ser calculada, pero sí Sª, con valor 71.

X	f
-3	9
-2	7
-1	4
0	10
1	5
2	5
3	10
	50

Media: 0,00
Desviación tipo: 2,14
D. tipo acotada: 71
CV de Pearson: (sin solución)

En definitiva, vamos a utilizar Sª como medida de acotación de la desviación tipo, tanto para interpretar una dispersión, como para comparar dispersiones de conjuntos distintos de una variable. No obstante, es bueno conocer la CV de Pearson, puesto que es el índice de mayor uso en este cometido.

Algo sobre sentido común y cálculos con tablas

Hemos abordado el cálculo de la moda, la mediana, la media aritmética y sus respectivas medidas de bondad de representación, a partir de conjuntos de datos. Pero también hemos aprendido a utilizar tablas. En muchas ocasiones, manejamos tantos datos que no realizaremos cálculos directamente con ellos, sino con las tablas. No sería necesario hacer ningún tipo de aclaración si tenemos el sentido común bien ejercitado. Pero a veces aprendemos a no utilizarlo y, en su lugar, acudir a la memoria pura. Por si es tu caso, vamos a ver qué hacemos cuando los cálculos se hacen con tablas. Para no alargarnos en exceso, probemos el estado de nuestra lógica y sentido común para el cálculo de la media aritmética.

Imagina que encontramos un grupo de seis estudiantes que, al preguntarles, dicen que utilizan ocho redes sociales. ¿Cuántas redes sociales acumulan entre el grupo, sin distinguir si son iguales o distintas? Podría hacer una suma: 8+8+8+8+8+8=48, pero tal vez tú no resuelvas el problema así. Quizá razones del siguiente modo: 6 estudiantes, cada cual con 8 redes, entonces un total de 6x8=48 redes. ¡Bravo!

Esto, pensarás, lo hace cualquiera. Pues en mi experiencia comprueba que los estudiantes se olvidan de ello cuando van a calcular una media aritmética a partir de una tabla de frecuencias. Recuperemos una tabla de este mismo documento:

X_i	f_i	$\%_i$	F_i	$\%a_i$
1	20	25	20	25
2	32	40	52	65
3	16	20	68	85
4	8	10	76	95
5	4	5	80	100
Σ	80	100		

¿Cuál es la suma de datos?

Obviamente no es 1+2+3+4+5 = 15. Eso es la suma de valores. Tenemos n=80 datos. ¿Cuánto suman? Al hacer una pregunta similar, sea una suma, una media aritmética, una desviación tipo... a partir de una tabla, una respuesta muy extendida es sumar los valores de la columna. ¡Error! La respuesta correcta es hacer como en el caso de los 6 estudiantes y las 8 redes por estudiante. Leo la tabla: 20 datos con el valor 1 (20) + 32 datos con el valor 2 (64) + 16 datos con el valor 3 (48) + 8 datos con el valor 4 (32) + 4 datos con el valor 5 (20), total 184. En otras palabras, la suma de los datos es la suma de las multiplicaciones de cada valor por la frecuencia de veces que aparece (no es una regla mnemotécnica, es una lectura literal de lo que tú piensas que ya sabes).

Observa la fórmula de la media aritmética o de la desviación tipo. En el numerador se encuentra un sumatorio. Se refiere a la suma de los datos (los 80 datos, en este caso). Si vamos a realizar el cálculo con una tabla, no hace falta hacer 80-1 sumas, sino unas pocas sumas de unas pocas multiplicaciones, como acabamos de hacer. En ese caso, si X representa al valor en lugar de al dato, entonces, la fórmula para cálculos con una tabla puede reescribirse del siguiente modo:

$$\overline{X} = \frac{\sum X_i f_i}{n} \qquad S = \sqrt{\frac{\sum (X_i - \overline{X})^2 f_i}{n}}$$