

Acerca del sentido común y del extendido fenómeno *yonosedeanalisisdedatosismo*¹

Vicente Manzano-Arrondo, 2014

Motivación

Tras años de experiencia en la docencia de análisis de datos, he observado que los estudiantes tienden a generar algo así como una actitud de defensa ante cualquier cosa que parezca matemáticas. Esa actitud de autoprotección frente a contenidos asociados a lo más desagradable de la existencia humana, resulta contraproducente. Impide comprender el sentido de lo que se hace y, por tanto, aplicar los conceptos en situaciones nuevas.

El análisis de los datos es mucho sentido común, mucha comprensión y ... un momento que haga memoria ... ¿qué cosa más? ... ya casi lo tengo ... ¡Ah! ¡Sí! ... Algo de cálculo. Pero esto último es la parte más técnica, ante la que se puede contar con chuletas, formularios, guías breves y otras ayudas. Lo curioso es que buena parte del contenido del análisis de datos es potencialmente fácil de comprender pero prácticamente incomprendido.

He bautizado jocosamente y malintencionadamente con *yonosedeanalisisdedatosismo* al siguiente fenómeno: al encontrarse ante una situación que resolveríamos con soltura, naturalidad y sin error en la vida cotidiana, pero inserta en una clase de estadística, se activa algo así como “yo no sé de esto, he de buscar un recurso que me permita memorizar o automatizar el proceso para llegar al resultado correcto”. Este fenómeno se complementa con el siguiente matiz: “¿Qué querrá el profesor que yo haga? Es decir ¿qué es un resultado correcto?”

Este documento pretende ser ilustrativo. No es un tema oficial que, como es hábito, llegue a ser materia de preguntas en un examen. Por ello me tomo algunas licencias, como es acudir a un lenguaje más literario (aficionado) que académico (ídem). He escogido un asunto que me persigue en las clases: el significado de la frecuencia absoluta en las fórmulas de cálculo, con independencia de que sean sencillas o densas. Es una buena concreción de lo que estoy describiendo.

No sé si alguna vez has escuchado o visto el famoso chiste de la magdalena. Varios humoristas lo han llevado a escena. Un hombre entra en un bar. El camarero se acerca. El cliente le pide un café y una magdalena². El camarero le responde azorado que puede ponerle otro dulce o acompañante, incluyendo tostadas con todo tipo de ungüento, pero que no quedan madalenas. El cliente piensa unos instantes y le dice “entonces deme un zumo de naranja... y una madalena”. El

1 Aviso: este documento cuenta con diversas notas a pie de página. Todas son absolutamente prescindibles. El autor no se hace responsable de las consecuencias psicopatológicas que la lectura de tales notas puedan causar en quien las lea. Ante la duda y como es habitual en la firma de contratos, no lea usted la letra pequeña.

2 Una magdalena, también conocida como madalena, es una especie de bizcocho de menores dimensiones, cubierto parcialmente por un papel semi-hidrófobo al que se obstina en permanecer adherida. Apreciada por gentes de toda procedencia social, se la ve habitualmente en desayunos.

camarero vuelve a expresarle, con otro discurso, esta vez más extenso, el problema de la inexistencia. La reacción del cliente es volver a pensar y volver a pedir un líquido (pongamos por caso la habitual infusión de salvia fresca con jengibre) igualmente acompañado por una madalena. Así transcurre una y otra vez. La desesperación del camarero va en aumento. Acude a todo tipo de recursos creativos, a sugerencias, a peticiones desesperadas, todo ello teniendo como respuesta la invariable inmutabilidad del cliente. El chiste tiene diversos finales, incluyendo el asesinato, el suicidio y el paro cardíaco. Es, sin lugar a dudas, humor negro. Nos reímos porque el que se muere es otro. Aunque la situación estimula la carcajada, es pertinente reconocer la tragedia en la historia; una tragedia que no se comprende hasta que no se vive.

En pocas palabras, yo soy el camarero y algunos estudiantes son los clientes. La madalena es el significado de la frecuencia en una fórmula de cálculo. Hasta el momento no ha tenido lugar asesinato ni suicidio. Pero no puedo garantizar nada.

Tal vez tú no tengas problemas con esto. Pero si conoces a alguien que lo sufra, pásale el texto y probemos suerte. El fenómeno no es un indicador de algún déficit de inteligencia o intelectual. Lo hemos aprendido. Tal vez seamos víctimas de nuestra propia historia, de malas docencias previas, de imaginarios colectivos, de lo que sea. Cuando digo que observo que esto no se entiende, no estoy afirmando que la gente sea tonta. No lo es más que yo. Ahí no reside la explicación. La gente es, sencillamente, torpe. Y no pasa nada. Tomado al azar un individuo del planeta, comprobaremos que ignora más de lo que sabe y que es torpe en mayor número de tareas que aquellas que realiza con maestría. Así que bienvenidos y bienvenidas a la dimensión mortal de la gente normal³, donde también resido yo. Este documento es para esa gente, sin complejos, como decía un anuncio.

Sumando espero

Un buen día llego al aula sin ganas de dar clase. No sé en qué entretener al estudiantado. ¡Bien! ¡Se me ocurrió una idea! Voy a hacer que suméis hasta que pueda freírse un huevo en la coronilla⁴. Vais a sumar mil números. Ahí va:

Suma 800 veces el valor 3 y 200 veces el valor 5

Lo suelto. Me siento. Acaricio la idea del asueto. Enciendo el smartphone, dispuesto a practicar el deporte de evadirme del cerebro. Se me ocurre levantar la vista⁵ y os veo con cara de “Vale. Ya lo he hecho. Y ahora qué”. ¡Habéis sumado mil datos en menos de un minuto! Pregunto por el resultado... Alguien dice con desgana “tres mil cuatrocientos”. Así que rápido y correcto. Asombroso. Como dice la pertinente chirigota *Las verdades del banquero*, “no doy crédito”.

3 Si te has habituado a ver la típica película de Disney donde la normalidad es odiosa y el/la protagonista se comporta como si lo único que tuviera sentido en la vida es el éxito clamoroso, la admiración externa, el aplauso multitudinario y, en definitiva, la ausencia sobresaliente de anonimato, entonces la bienvenida con que te he recibido te resultará similar a una ofensa. Créeme, no era mi intención, en absoluto. Pero ya sabes, aquí en Cádiz... (y Sevilla está cerca). Y si no lo sabes, pregunta, que se adquiere cultura.

4 Coronilla es la zona de la cabeza donde podría mantenerse en equilibrio una corona de pequeñas dimensiones cuando la barbilla se encuentra aproximadamente a cinco centímetros del esternón y el individuo propietario del cuerpo se encuentra en posición vertical. (Nota del editor). Barbilla es la zona del rostro en donde podría colgar una barba de pequeñas dimensiones, caso de existir. Si existe, los cinco centímetros de distancia entre la barbilla y el esternón necesarios para que la coronilla se encuentre en la cota más alta de la estructura, desaparecen. Utilícese como criterio cuando los pelos de la barbilla hacen cosquillas.

5 Obvia decirlo, pero por si estás navegando mientras lees esto y por tanto tu nivel de atención está dividido, si he levantado la vista es que todavía no terminó de encenderse el smartphone.

La explicación de la tarea es más larga que su ejecución. Pero probemos:

Obviamente no voy a sumar 3 y 3 y 3 y 3 y así 800 veces. Eso es ridículo. Como el 3 está 800 veces, eso dará 3 por 800 igual a 2400. Lo mismo con lo otro, que es más fácil: el 5 repetido 200 veces es 1000. Sumo los dos y me da 3400. Eso es todo.

Simbolizando

Vale. Asumo el golpe⁶. Voy a tener que trabajar.

Bien. Entonces tenemos que hay mil datos pero solo dos valores, cada uno de los cuales se repite un número de veces diferente. Si dispongo espacialmente las operaciones, lo que he hecho ha sido:

$$\begin{array}{rcl} 3 & \times & 800 = 2400 \\ 5 & \times & 200 = 1000 \\ & & \text{total} = 3400 \end{array}$$

Ya sabemos que el número de veces que ocurre algo es la frecuencia. Así lo llamamos tanto en el bar de la esquina como en clase. Cuando alguien va a la consulta de medicina porque le duele la cabeza de vez en cuando y está comenzando a preocuparse, no es raro que el profesional le pregunte “¿Con qué frecuencia le ocurre?”. Y le entendemos perfectamente sin suponer que estamos en una clase de estadística. Es un término de nuestro vocabulario, que utilizamos con soltura. Así que alguien calcula 2400 como resultado de multiplicar el valor 3 por el número de veces que aparece, es decir, por su frecuencia. Como sabes, los símbolos que utilizamos para valor y frecuencia son, respectivamente, X_i y f_i .

Con tales símbolos, podemos reescribir las operaciones que habéis hecho, así:

X_i	f_i	$X_i f_i$
3	800	2400
5	200	<u>1000</u>
		3400

Parece que hemos complicado el asunto. Tal vez estés en lo cierto. Pero no estamos haciendo operaciones nuevas. Lo que ocurre es que estamos poniendo símbolos a lo que sabemos hacer perfectamente: sumar cosas que se repiten. Las cosas son los valores numéricos y las frecuencias indican cuánto se repiten. Ya está.

Pero no. No está. Ahora aparece el fenómeno. Os voy a entretener con algo ligeramente más elaborado. Voy a pedir que calculéis la media aritmética de la siguiente variable. La tabla de frecuencias es la siguiente:

6 Reconozco incluso que no me ha dolido. O todavía mejor expresado: sarna con gusto no pica, célebre dicho español que viene a decir: si aprecias más lo positivo del resultado que lo negativo del proceso, bienvenido sea el proceso negativo. Me diste un golpe, pero ha sido por una buena causa. Si pudiera poner una nota a pie para esta nota a pie tendría que mencionar mi desacuerdo con la máxima de Maquiavelo, puesta repetidas veces en boca de Fernando el Católico: el fin justifica los medios. A eso no me refiero con la sarna y el picor. Pero vamos a dejarlo aquí. Sigue tú cuando quieras.

X_i	f_i
2	9
4	11
6	12
7	8
11	<u>10</u>
	50

Pues allá vamos. ¿Cuál es la media?

La media aritmética se obtiene dividiendo la suma de datos entre el número de datos. No vamos a disertar aquí acerca de su significado⁷. Vayamos únicamente a su cálculo. Hemos dicho *suma de los datos*. Pues bien, no sé si te pasará a ti, pero algunos estudiantes hacen lo siguiente:

$$2 + 4 + 6 + 7 + 11 = 30$$

Eso no es la suma de los datos, sino la suma de *algunos* datos, los *diferentes entre sí*. Contamos con 50, no con 5. Hay 9 datos con el valor 2, 11 con el valor 4, etc. Los hemos despreciado a casi todos. Esa operación, consistente en sumar solo los datos con valores diferentes, deja fuera a 45 de los 50. Aunque sea una tabla de frecuencias, la lógica sigue siendo la misma del ejemplo de los mil datos. Vamos a ver extendidos a estos 50, sin estar dispuestos en una tabla de frecuencias. Y los ordeno según su cuantía, para facilitar la conclusión:

2 2 2 2 2 2 2 2 2 2 4 4 4 4 4 4 4 4 4 4 6 6 6 6 6 6 6
6 6 6 6 6 7 7 7 7 7 7 7 7 11 11 11 11 11 11 11 11 11 11

Si te digo que sumes esos datos, no creo que hagas esto:

$$2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 4 + 4 + 4 + \dots$$

Más bien, para ahorrar esfuerzo, dirás: el 2 está 9 veces, así que $2 \times 9 = 18$, a lo que he de sumar los once 4, es decir $18 + (4 \times 11) = 18 + 44 = 62$, a lo que he de sumar los doce 6, es decir... Si en lugar de tener el listado continuo de los 50 datos los disponemos en la tabla de frecuencias para facilitar los cálculos, lo que ocurre es precisamente que los cálculos se facilitan, no que se evitan:

X_i	f_i	$X_i f_i$
2	9	18
4	11	44
6	12	72
7	8	56
11	<u>10</u>	<u>110</u>
	50	300

Dado que la media es la suma de datos entre el número de datos, el resultado es:

⁷ Como ya has observado, este documento no tiene absolutamente nada irrelevante. Todo el texto es importante para descubrir el significado de la existencia humana a través de las estrategias de análisis de datos. Kevin Costner, creo recordar que en *Tin Cup*, afirma que el béisbol es la esencia del misterio de la vida en el universo. Si él lo dice, ¿por qué no ser el análisis de los datos o cualquier otra cosa parecida?

$$\bar{X} = \frac{\sum_{i=1}^{50} X_i}{n} = \frac{\sum X_i}{n} = \frac{2+2+2+2+\dots}{50} = \frac{300}{50} = 6$$

Como observarás, en el tercer elemento de la igualdad múltiple, he eliminado las especificaciones del sumatorio. En lugar de indicar “Sumatorio desde $i=1$ a 50”, sencillamente pone “Sumatorio”. Esto es un convenio para simplificar la expresión. Pero hemos de saber qué estamos haciendo. En la práctica, yo sé que no voy a hacer literalmente lo que dice la fórmula, es decir, no voy a sumar 50 datos directamente, sino que utilizo un atajo, el que ya conocemos, multiplicando los valores por sus frecuencias. Para que la fórmula exprese literalmente el aprovechamiento de las frecuencias, deberíamos hacerle alguna modificación:

$$\bar{X} = \frac{\sum_{i=1}^5 X_i f_i}{n} = \frac{\sum X_i f_i}{n} = \frac{18+44+\dots}{50} = \frac{300}{50} = 6$$

Esto de andar con diferentes subíndices y de estar escribiéndolos continuamente es todo un trabajo. Nos lo podemos ahorrar. Las expresiones de cálculo pueden ser muy sencillas. Pero para simplificarlas es necesario que no perdamos el norte, es decir, que no se nos olvide qué es lo que estamos haciendo. En otras palabras y si el número total de datos se representa por n mientras que el número de valores diferentes es k :

Sé que $\sum X_i$ expresa “suma de los n datos”, bien sea literalmente ($\sum_{i=1}^n X_i$) o mediante un atajo ($\sum_{i=1}^k X_i f_i$)

Varianciando⁸

Muchas personas en clase no hacen nunca eso de sumar solo los valores diferentes cuando lo que necesitan es sumar todos y cada uno de los datos⁹. Calculan perfectamente la media aritmética, sea con el listado original o a partir de una tabla de frecuencias. También es cierto que buena parte de ellos no sigue con esa lógica cuando la fórmula resulta menos familiar. Vamos a verlo con el cálculo de la varianza, tanto en su expresión original como en la de cálculo simplificado.

El problema que abordamos ahora es calcular la varianza de los datos del ejemplo anterior. Sabemos que la varianza es una media de distancias cuadráticas. Repetimos, más despacio:

8 Permítaseme seguir neologizando sinvergonzosamente, pues dado que este asunto no será objeto de pregunta en examen alguno, bien vale la licencia de operar con el lenguaje al antojo. Este pie de página tiene igualmente el interés de retar al traductor.

9 Dejo libre la interpretación cuantitativa del término “muchas”. Lo he utilizado especialmente para facilitar que si le pasa a alguien, pueda sentirse participe de un grupo numeroso. Gracias a la propensión que parece tenemos para concebirnos parte de la mayoría (desde una estética de la minoría), los anuncios venden que da miedo. Pero ya conoces el dicho: dos son compañía y tres, multitud. De hecho, cuando alguien se excede, se le reprocha con “te has pasado tres pueblos”. Así pues, “muchas”, como muchos otros muchos, es algo mucho relativo. Vale, lo dejo, por si ya te parece mucho. Yo he conseguido mi objetivo: que la tabla pase a la siguiente página.

- Distancia: $X_i - \bar{X}$
- Distancia cuadrática: $(X_i - \bar{X})^2$
- Suma de distancias cuadráticas: $\sum (X_i - \bar{X})^2$
- Media de distancias cuadráticas: $\frac{\sum (X_i - \bar{X})^2}{n}$

Vamos a utilizar la tabla de frecuencias para incluir lo que necesitamos. Además de las columnas habituales de valores (X) y frecuencias (f), vamos a añadir la de distancias a la media (d) y el resultado de elevarlas al cuadrado (d2):

X	f	d	d2
2	9	-4	16
4	11	-2	4
6	12	0	0
7	8	1	1
11	<u>10</u>	5	25
	50		

Como cabría esperar, muchas personas (más que en los casos anteriores), suman las distancias cuadráticas del siguiente modo:

$$16 + 4 + 0 + 1 + 25 = 46$$

Es un error, como posiblemente ya tienes claro. Es un error porque hay ¡50 distancias cuadráticas! Aunque solo 5 de ellas son diferentes. La distancia -4 se encuentra en 9 ocasiones, ya que hay 9 datos con valor X=2, es decir, alejados de la media en cuatro unidades por debajo. Es decir, hay 9 distancias cuadráticas con valor 16. Si incluimos esta información (d2f) en la tabla:

X	f	d	d2	d2f
2	9	-4	16	144
4	11	-2	4	44
6	12	0	0	0
7	8	1	1	8
11	<u>10</u>	5	25	<u>250</u>
	50			446

La suma de las 50 distancias cuadráticas es 446, no 46. Hay un ligero error de cuantía 400 (un error del 870%¹⁰). En términos simbólicos:

$$S^2 = \frac{\sum_{i=1}^{50} (X_i - \bar{X})^2}{n} = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{16+16+16+\dots}{50} = \frac{446}{50} = 8,92$$

¹⁰ No somos un partido político al uso haciendo promesas en unas elecciones, así que un error del 870% no es despreciable.

o bien:

$$S^2 = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2 f_i}{n} = \frac{\sum (X_i - \bar{X})^2 f_i}{n} = \frac{144+44+\dots}{50} = \frac{446}{50} = 8,92$$

Como en el caso de la media aritmética, es indiferente que yo utilice la primera expresión para la fórmula y que calcule según la segunda. El numerador no varía. En otras palabras:

Sé que $\sum (X_i - \bar{X})^2$ expresa “suma de las n distancias cuadráticas”, bien sea literalmente ($\sum_{i=1}^n (X_i - \bar{X})^2$) o mediante un atajo ($\sum_{i=1}^k (X_i - \bar{X})^2 f_i$)

Simplivariando¹¹

El cálculo del índice de variación que conocemos asombrosamente como *varianza* puede ser simplificado. La fórmula anterior es la que se ajusta literalmente a la idea, es decir, a expresarse como media de distancias cuadráticas. No obstante, andar restando continuamente y elevando el resultado al cuadrado, especialmente cuando la media aritmética te obliga a arrastrar decimales, es un fastidio digno de mención. Hay una estrategia más sencilla, que tal vez ya conozcas y que surge de resolver el paréntesis de la distancia cuadrática. El resultado es:

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

Leyendo casi con estilo poético: la media de las distancias cuadráticas es lo mismo que la media de los cuadrados menos el cuadrado de la media. Observa que volvemos a tener una suma. No hay forma de librarse de ella. Si hay suma, hay confusión en potencia, pues una vez más ¿qué hacer con las frecuencias? La respuesta es sencilla: lo mismo de siempre. Pero no nos vale con algo tan general ¿verdad? Vamos a ello. Comenzamos con la tabla, donde he añadido el cuadrado de cada valor (X^2):

X	f	X ²
2	9	4
4	11	16
6	12	36
7	8	49
11	<u>10</u>	121
	50	

Ya sabes la historia: la suma de todos los cuadrados de los números es la suma de los 50 cuadrados, no la de los 5 diferentes. No es esto:

$$4 + 16 + 36 + 49 + 121 = 226$$

¹¹ Ya te queda na. Quién sabe si cuando superes esta asignatura recordarás más el chiste de la madalena o los neologismos que el cálculo de MAD. Se admiten apuestas.

Si no esto:

$$4 \cdot 9 + 16 \cdot 11 + 36 \cdot 12 + 49 \cdot 8 + 121 \cdot 10 = 2246$$

Luego

$$S^2 = \frac{\sum X_i^2}{n} - \bar{X}^2 = \frac{4+4+4+\dots}{50} - 6^2 = \frac{36+176+\dots}{50} - 6^2 = 8,92$$

Y, en definitiva:

Sé que $\sum X_i^2$ expresa “suma de los n cuadrados”, bien sea literalmente ($\sum_{i=1}^n X_i^2$) o mediante un atajo ($\sum_{i=1}^k X_i^2 f_i$)

Madificinger Z¹²

Si se me ocurre relatar todo lo anterior en clase, sea en la misma sesión o troceada por entregas semanales en vivo, un efecto *frecuente* es la profusión de indicadores de hastío, con caras de “ya estamos otra vez con la misma canción”.

Una de mis piezas favoritas es *Variaciones y fuga sobre un tema de Purcell Op. 34* de Benjamín Britten¹³. Doy clases de análisis de datos, donde el concepto de variabilidad es fundamental. Así que cuando varío explicaciones sobre un mismo tema me viene a la cabeza una tontería: *efecto opus 34*. En ningún momento el efecto op34 es más evidente que en el asunto de las frecuencias, muy especialmente aplicado al epígrafe de las distancias estandarizadas.

Como sabes, una distancia estandarizada es precisamente eso: una distancia que se ha estandarizado, usualmente simbolizada con la letra Z. Más despacio:

- Distancia: $X_i - \bar{X}$
- Distancia estandarizada: $\frac{X_i - \bar{X}}{S}$

He indicado ya en otro lugar la inestimada utilidad de medir el precio de las cosas mediante un objeto de consumo con significado. En mis tiempos de estudiante medía todo lo mensurable cuantitativamente con bocadillos de chorizo pamplonica. Gracias a ser especialmente asequible,

12 Mi hermano Carlos es una de estas personas que te comen la moral. Para quien no entienda la expresión, me refiero a un fenómeno positivo que te hace sentir una mosca insignificante. Es sordo casi por completo. Sin embargo y entre otras habilidades, toca el piano *de oído* y lo hace con una maestría sorprendente. En la misma línea, a los seis años realizó una pintura a trazo de lápiz sobre Mazinger Z, los dibujos animados que arrasaban por aquel entonces. El dibujo dejó alucinados a todos los miembros de la familia y a parte del extranjero. Ahora tiene 41 años y se dedica al diseño a mano y por ordenador de todo tipo de escenarios. Eso significa que han pasado 35 primaveras desde la época en que Mazinger Z era el referente de miles de niños. Hemos mejorado sensiblemente. En lugar de un robot gigante que hace justicia destrozando edificios mientras lucha con robots enemigos, ahora contamos con Los Simpsons.

13 Por la razón que sea, cuando alguien habla de que le gusta una obra de música clásica (llamada genéricamente así aunque sea barroca, renacentista o romántica, por ejemplo), es tachada de pedante. No ocurre lo mismo si pongo como ejemplo otras piezas que me encantan, como la versión de Elvis Presley de *Blue Suede Shoes*, *Rasputín* de Boney M, *Como el agua* de Camarón, *What's my name* de Rihanna o el impresionante *El Garrotín* de Smash.

constituía la base de mi dieta. Sobreviví y hoy puedo estar escribiendo esta obra fundamental. Si alguien me decía el precio de un calzado, de un auto, del alquiler de un piso, de un ordenador... lo que yo hacía para interpretar esa cuantía era dividirla por el precio de mi bocadillo. De este modo podía concluir cosas como: esos zapatos valen 35 bocadillos de chorizo pamplonica (BCP); ese alquiler, 200 BCP; ese ordenador es una ganga, por solo 89 BCP; etc. Si voy al extranjero y quiero tener una idea de lo que me cuesta la vida, pregunto por lo más parecido a un BCP. La distancia estandarizada tiene el mismo significado. Ya lo sabes, pero te lo recuerdo. En lugar de expresar la distancia que separa un valor respecto a lo característico de su conjunto de datos, sin más, lo que hacemos es re-expresarla en número de desviaciones, de tal forma que a mayor dispersión general, la distancia particular es menos relevante. Como ya hemos abordado esto, no me alargó más.

Lo interesante aquí es la propiedad “La suma de las Z da 0 como resultado”. Es un efecto deducido de su concepto, pues resulta intuitivo que si la media se encuentra en el centro de las cuantías, si sumamos las diferencias positivas a la media (valores situados por encima) y restamos las negativas (valores situados por debajo), el resultado debe ser cero. Además de intuitivo, se demuestra aritméticamente.

Así que, tras las explicaciones sobre el papel de la frecuencia en los cálculos abreviados para medidas estadísticas, llega el momento de la prueba de fuego: comprobar que, en efecto, la suma de las distancias estandarizadas da como resultado un cero redondo. En ocasiones es un poco ovalado e incluso cuadrado, depende de la agresividad de redondeo que se haya aplicado en los cálculos. Pero si no da exactamente cero, debe obtenerse razonable y aproximadamente cero. He aquí un ejemplo:

X	f	Z
1	13	-0,83
2	10	-0,55
3	9	-0,28
4	7	0,00
7	4	0,83
12	<u>7</u>	2,22
	50	

La media y desviación tipo de este conjunto de datos tienen como valores respectivos 4 y 3,61. Como ya no tengo nada en lo que entretenerme, os pido comprobar que el resultado de sumar las distancias estandarizadas es en efecto cero. Aquí es cuando alguien me dice “pues a mí me da 1,39”. No sé cómo me contengo.

Tras escucharlo, me invade con toda su crudeza la tragedia del camarero y la madalena. Recuerdo la desesperación del pobre obrero del mundo de la restauración. Es difícil superar la sospecha de que el cruento final puede ocurrir en el aula. Sinceramente prefiero el asesinato al suicidio, es decir que le toque a otro. Espero que me queden todavía muchos cálculos basados en BCP. Pero me dan ganas de gritar “¡Alma de cántaro¹⁴! ¿Otra vez?” Las mismas caras que antes mostraban hastío ahora expresan “¿Qué le pasa a este ahora? ¿Qué cosa he dicho?”

La única forma de obtener 1,39 o algo similar es sumar únicamente los **seis** valores diferentes que se observan en el conjunto de las **cincuenta** distancias estandarizadas. La suma de

14 Expresión de origen incierto, contenida ya en Don Quijote y popularizada por el personaje Pelayo (José Antonio Sayagués) en la serie *Amar en tiempos revueltos*. Significa inocente, cándido y, con un poco más de carga emotiva, tonto. En femenino, tonta o cándida, pero no *alma de cántara*.

las distancias, como en todos los casos anteriores, se refiere a *todas* ellas y no solo a una élite, seleccionada por el criterio disneyano¹⁵ de parecer diferente destacando del resto.

Lo correcto no es:

$$- 0,83 - 0,55 - 0,28 + 0,83 + 2,22 = 1,39$$

Para encontrar la suma, va a resultar muy útil añadir a la tabla una nueva columna (Zf) que expresa el resultado de sumar el valor Z un total de f veces:

X	f	Z	Zf
1	13	-0,83	-10,79
2	10	-0,55	-5,50
3	9	-0,28	-2,52
4	7	0,00	0,00
7	4	0,83	3,32
12	<u>7</u>	2,22	<u>15,54</u>
	50		0,05

Dado que hemos redondeado en el cálculo de la desviación tipo y de las distancias estandarizadas, el resultado tiene un error de 5 centésimas, que a los efectos de lo que estamos abordando, no es relevante. De hecho, utilizando tres cifras decimales, la distancia estandarizada que se corresponde con el valor X=2 es Z=-0,554. En tal caso, la suma de las 10 distancias con valor Z=-0,554 será -5,54, con lo que el resultado final tiene únicamente un error de una centésima. Para caminar con más seguridad, sería recomendable que los elementos de cálculo cuenten con no menos precisión que el resultado final. Si la Z se ha calculado con dos cifras decimales, al multiplicarlas por las frecuencias, que rondan el valor 10, el resultado transforma en precaria a la segunda cifra decimal resultante y solo habría que considerar una. Pero esto es ya otra historia.

En definitiva, en cualquier fórmula y para terminar:

Sé que $\sum Z_i$ expresa “suma de las n distancias estandarizadas”, bien sea literalmente ($\sum_{i=1}^n Z_i$) o mediante un atajo ($\sum_{i=1}^k Z_i f_i$)

Y con esto terminamos. Sé feliz¹⁶.

15 Para evitar confusiones, aclaro que no me refiero al lugar del dorso del cuerpo situado entre espalda y muslos en referencia al genial cineasta estadounidense. Vuelve a ser un neologismo que afecta de nuevo a la tendencia a destacar, expresada magistralmente por María Isabel López en *Antes muerta que sencilla*, dentro del álbum *No me toques las palmas que me conozco*.

16 Obviamente es una sugerencia y la expresión de un deseo. No te imagino un día visiblemente alegre ante alguien que te pregunta “Oye ¿y tú por qué eres tan feliz?”, respondiéndole “Nunca se me había ocurrido. Pero me lo dijo mi profe de análisis de datos”. Ahora que lo pienso, no está mal para iniciar el guión de una película.