Ejercicio completo argumentado para los temas “Conocer una variable” e “Interpretar un caso”

Hemos preguntado a un grupo de personas sobre el número de veces que han desayunado una bebida excitante (té, chocolate, café) en las dos últimas semanas. Los resultados son estos:

4	2	7	2	2	3	5	4
2	5	2	4	3	4	3	5
4	2	5	3	4	2	3	7
2	7	3	10	2	8	4	2
4	7	4	3	4	3	2	8

- 1) Construye una tabla de frecuencias con solo tres columnas: valores, frecuencias absolutas y distancias estandarizadas.
- 2) Comprueba que la suma de las distancias estandarizadas da 0 como resultado.
- 3) Indica cuál es el valor de las siguientes medidas de posición: Q_1 , Q_3 , D_5 , D_8 , P_{37} , P_{54} y P_{85} .

Solución

Para resolver el primer problema, necesitamos contar con los valores de dos estadísticos: la media aritmética y la desviación tipo, pues ambos son necesarios para calcular las distancias estandarizadas. Vamos a llevar a cabo estos cálculos a partir de los 40 datos originales y, después, a partir de la tabla de frecuencias.

El valor de la media aritmética, a partir de los 40 datos, es:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{4+2+7+2+\dots+4+3+2+8}{40} = \frac{160}{40} = 4$$

Si lo calculamos con la tabla de frecuencias, lo que hacemos es ahorrar sumas múltiples. Así, resulta que el valor 2 se encuentra presente en 11 ocasiones, es decir, hay 11 datos con el valor 2, por lo que en lugar de sumar el 2 once veces, lo que hacemos es sumar su multiplicación $2 \cdot 11 = 22$.

X_i	f_i	$X_i f_i$
2	11	22
3	8	24
4	10	40
5	4	20
7	4	28
8	2	16
10	1	10
	40	160

La fórmula es la misma, pero podemos hacer una pequeña modificación para señalar que en lugar de simbolizar con X_i a los 40 datos, vamos a utilizar X_i para simbolizar a los 7 valores:

$$\bar{X} = \frac{\sum (X_i f_i)}{n} = \frac{2 \cdot 11 + 3 \cdot 8 + \dots + 8 \cdot 2 + 10 \cdot 1}{40} = \frac{160}{40} = 4$$

La desviación tipo es la raíz cuadrada de la media de las distancias cuadráticas, según se lee en su fórmula de cálculo:

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

Así pues, para obtenerla hay que calcular las diferencias con la media aritmética:

0	-2	3	-2	-2	-1	1	0
-2	1	-2	0	-1	0	-1	1
0	-2	1	-1	0	-2	-1	3
-2	3	-1	6	-2	4	0	-2
0	3	0	-1	0	-1	-2	4

Elevarlas al cuadrado:

0	4	9	4	4	1	1	0
4	1	4	0	1	0	1	1
0	4	1	1	0	4	1	9
4	9	1	36	4	16	0	4
0	9	0	1	0	1	4	16

Y resolver la fórmula:

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} = \sqrt{\frac{0+4+9+4+\dots+0+1+4+16}{40}} = 2$$

Como hemos hecho en el caso del cálculo de la media aritmética, podemos acudir a la tabla de frecuencia, operando con los 7 valores en lugar de con los 40 datos, ya que del mismo modo que el valor 2 se encuentra 11 veces, su distancia a la media aritmética (2-4= -2) está presente también 11 veces y, por tanto, su distancia al cuadrado (-2²= 4) está también 11 veces. Así que (simbolizando con d2 a la distancia cuadrática, es decir, al cuadrado de la diferencia de cada valor respecto a la media):

Xi	fi	d2	d2fi
2	11	4	44,00
3	8	1	8,00
4	10	0	0,00
5	4	1	4,00
7	4	9	36,00
8	2	16	32,00
10	1	36	36,00
	40		160,00

$$S = \sqrt{\frac{\sum (X_i - \bar{X})^2 f_i}{n}} = \sqrt{\frac{44+8+\dots+32+36}{40}} = 2$$

Ya que tenemos el valor de la media aritmética y de la desviación tipo, podemos calcular las distancias estandarizadas para cada uno de los 40 datos:

0	-1	1,5	-1	-1	-0,5	0,5	0
-1	0,5	-1	0	-0,5	0	-0,5	0,5
0	-1	0,5	-0,5	0	-1	-0,5	1,5
-1	1,5	-0,5	3	-1	2	0	-1
0	1,5	0	-0,5	0	-0,5	-1	2

O bien, mediante la tabla de frecuencias:

X_i	f_i	Z_i
2	11	-1,00
3	8	-0,50
4	10	0,00
5	4	0,50
7	4	1,50
8	2	2,00
10	1	3,00
	40	

En cualquiera de ambos casos, hemos utilizado la misma estandarización:

$$Z = \frac{X_i - \bar{X}}{S} \quad \frac{2 - 4}{2} = -1 \quad \frac{3 - 4}{2} = -0,5 \quad \dots \quad \frac{10 - 4}{2} = 3$$

Ya hemos respondido a la primera pregunta del problema: construir una tabla de frecuencias con las columnas de valores, frecuencias absolutas y distancias estandarizadas. Ahora, para la segunda, hemos de sumar las distancias y comprobar que el resultado es 0. Un primer impulso es sumar directamente desde la tabla. Si lo haces, observa lo que ocurre:

X_i	f_i	Z_i
2	11	-1,00
3	8	-0,50
4	10	0,00
5	4	0,50
7	4	1,50
8	2	2,00
10	1	3,00
	40	5,50

No hemos obtenido 0 como resultado, sino 5,5. ¿Por qué? Pues porque no hemos sumado *todas* las distancias estandarizadas, es decir, las 40, sino únicamente las 7 distintas que se consideran en la tabla. Si te molestas en sumar las 40 distancias estandarizadas, verás que así obtienes el resultado correcto:

$$\sum Z_i = 0 - 1 + 1,5 - 1 + \dots + 0 - 0,5 - 1 + 2 = 0$$

Con la misma lógica que hemos llevado hasta el momento, en lugar de sumar 11 veces la distancia $Z = -1$ (que se corresponde con los 11 datos que tienen el valor 2), podemos acudir a la tabla, pero haciendo los cálculos adecuados:

X_i	f_i	Z_i	$Z_i f_i$
2	11	-1,00	-11,00
3	8	-0,50	-4,00
4	10	0,00	0,00
5	4	0,50	2,00
7	4	1,50	6,00
8	2	2,00	4,00
10	1	3,00	3,00
	40		0,00

Finalmente, el problema pregunta por los valores que se corresponden con algunas medidas de posición. Para responder, nos vendrá muy bien disponer de una tabla con los porcentajes acumulados, ya que hemos decidido utilizarlos como referente para los percentiles. Del mismo modo, para no facilitar confusiones, será preferible traducir todo a percentiles. Así pues:

$$Q_1, Q_3, D_5, D_8, P_{37}, P_{54} \text{ y } P_{85} = P_{25}, P_{75}, P_{50}, P_{80}, P_{37}, P_{54} \text{ y } P_{85}$$

Ya que estamos, vamos a ordenarlos:

$$P_{25}, P_{37}, P_{50}, P_{54}, P_{75}, P_{80} \text{ y } P_{85}$$

Es más, para facilitar la locación de las medidas de posición (o puntos de corte del conjunto ordenado de datos), vamos a crear otra columna específica que indique qué puntos de corte se corresponden con cada valor, es decir, qué intervalos de percentiles (columna P_i):

X_i	f_i	%ai	P_i
2	11	28	1 – 28
3	8	48	29 – 48
4	10	73	49 – 73
5	4	83	74 – 83
7	4	93	84 – 93
8	2	98	94 – 98
10	1	100	99
	40		

Observa que el valor 10, que cuenta con una frecuencia $f_i=1$, no expande sus percentiles desde 99 a 100. Esto es así porque los percentiles se refieren a 99 puntos de corte que dan como resultado 100 partes. Luego, no tiene sentido el percentil 100, ya que no existe el punto de corte número 100.

- P_{25} se encuentra en el intervalo 1 a 28, luego, su valor es $X_i = 2$.
- P_{37} se encuentra en el intervalo 29 a 48, luego, su valor es $X_i = 3$.
- P_{50} se encuentra en el intervalo 49 a 73, luego, su valor es $X_i = 4$.
- P_{54} se encuentra en el intervalo 49 a 73, luego, su valor es $X_i = 4$.
- P_{75} se encuentra en el intervalo 74 a 83, luego, su valor es $X_i = 5$.
- P_{80} se encuentra en el intervalo 74 a 83, luego, su valor es $X_i = 5$.
- P_{85} se encuentra en el intervalo 84 a 93, luego, su valor es $X_i = 7$.