Módulo de autoaprendizaje Nº14
Tema: Varianza y desviación típica.

Objetivo: Comprender la varianza y desviación típica.

Definición:

Sin lugar a dudas la medida más usada para estimar la dispersión de los datos es la desviación típica. Esta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Al igual que la desviación media, está basada en un valor promedio de las desviaciones respecto a la media. En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final. Se define entonces la varianza de una muestra con datos repetidos como

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1}.$$

Evidentemente la varianza no tiene las mismas unidades que los datos de la muestra. Para conseguir las mismas unidades se define la desviación típica (algunas veces llamada desviación estándar) como la raíz cuadrada de la varianza

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1}}.$$

En el caso de que los datos no se repitan, estas definiciones se simplifican a

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} \quad ; \quad s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}.$$

En muchas ocasiones se definen varianza y desviación típica utilizando N en vez de N − 1 en el denominador, representando entonces la varianza una verdadera media aritmética del cuadrado de las desviaciones. Está claro que ambas definiciones llevan a valores muy parecidos cuando N es grande. El motivo de haber optado aquí por la definición con N − 1 es que ´esta da una mejor estimación de la dispersión de los datos. Téngase en cuenta que como la suma de las desviaciones xi − $\bar{x}$ es siempre, la desviaci´on del ´ultimo dato puede calcularse una vez que se conozcan las N − 1 anteriores. Es decir, solo se tienen N − 1 desviaciones independientes (se dice que el sistema tiene N − 1 grados de libertad) y se promedia entonces dividiendo por N −1, ya que no tiene mucho sentido promediar N números no independientes. Notes 'e además que cuando solo se tiene un dato (N = 1), en el caso de la definición con N en el denominador se obtendría una varianza 0, que no tiene mucho sentido, mientras que en la definición con N − 1 la varianza estaría indeterminada. En cualquier caso, siempre se puede obtener una desviación típica a partir de la otra multiplicando (o dividiendo) por $\sqrt{(N-1)/N}$

$$\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N}} = \sqrt{\frac{N-1}{N}}\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1}}.$$

La expresión $s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1}$. no es muy cómoda para calcular la desviación típica de forma rápida. A efectos prácticos, dicha expresión se puede transformar en otra más fácil de aplicar

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1} = \frac{\sum x_i^2 n_i - 2\sum x_i \overline{x} n_i + \sum \overline{x}^2 n_i}{N-1} =$$

$$= \frac{\sum x_i^2 n_i - 2\overline{x}\sum x_i n_i + N\overline{x}^2}{N-1},$$

donde se ha usado que $\sum_{i=1}^{k} n_i = N.$ Utilizando ahora la expresión para la media

$$s^2 = \frac{\sum x_i^2 n_i - 2\frac{1}{N}\sum x_i n_i \sum x_i n_i + \frac{N}{N^2}(\sum x_i n_i)^2}{N-1} = \frac{\sum_{i=1}^{k} x_i^2 n_i - \frac{1}{N}(\sum_{i=1}^{k} x_i n_i)^2}{N-1}.$$

La expresión anterior es más fácil de aplicar ya que bastar ‘a con calcular los sumatorios de los datos al cuadrado y de los datos, habiéndose calculado ya este ´ultimo para la media.

En cuanto a las propiedades de la desviación típica, es fácil ver que ´esta ser ‘a siempre positiva y sólo tendrá un valor nulo cuando todas las observaciones coincidan con el valor de la media. Además, si se define la desviación cuadrática respecto a un promedio a como

$$D^2 = \frac{\sum_{i=1}^{k}(x_i - a)^2 n_i}{N-1}.$$

Se puede demostrar que dicha desviación cuadrática será mínima cuando a = x. Es decir, la varianza (y, por tanto, la desviación típica) es la mínima desviación cuadrática. Para demostrarlo derivamos la expresión anterior respecto a a, e igualamos la derivada a 0 (condición necesaria para que D^2 sea mínimo)

$$\frac{\partial D^2}{\partial a} = 0 = \frac{-2\sum(x_i - a)n_i}{N-1}$$

$$\Rightarrow \quad \sum(x_i - a)n_i = 0 \quad \Rightarrow \quad \sum x_i n_i - a\sum n_i = 0$$

$$\Rightarrow \quad \sum x_i n_i - aN = 0 \quad \Rightarrow \quad a = \frac{\sum x_i n_i}{N} = \overline{x},$$

como queríamos demostrar. Esta propiedad le da además más sentido a la definición de la desviación típica.
Hay que indicar que la desviación típica no es una medida robusta de la dispersión. El hecho de que se calcule evaluando los cuadrados de las desviaciones hace que sea muy sensible a observaciones extremas, bastante más que la desviación media (dado que aparece un cuadrado). En definitiva, la desviación típica no es una buena medida de dispersión cuando se tiene algún dato muy alejado de la media. El rango intercuartílico nos daría en ese caso una idea más aproximada de cuál es la dispersión de los datos. El que la desviación típica sea la medida de dispersión más común se debe a su ´íntima conexión con la distribución normal, como se verá en sucesivos capítulos.
En la discusión sobre la media aritmética se vio como su cálculo se podía simplificar a veces si se realizaba una transformación lineal de la variable x a una nueva variable y, definida. En este caso, existe una relación muy sencilla entre las desviaciones típicas (sy y sx) de ambas variables, ya que

$$s_y = \sqrt{\frac{\sum(y_i - \overline{y})^2}{N-1}} = \sqrt{\frac{\sum(a + bx_i - a - b\overline{x})^2}{N-1}} = \sqrt{\frac{b^2\sum(x_i - \overline{x})^2}{N-1}} = bs_x.$$

De esta forma, una vez calculada la desviación típica de y, se puede evaluar la de x haciendo

$$s_x = \frac{s_y}{b}.$$

Se demuestra así además que, aunque la desviación típica depende de las unidades elegidas (a través de b), es independiente de un cambio de origen (dado por a).

Ejemplo:

a.

En el caso de una variable discreta

x_i	n_i	$x_i \times n_i$	$x_i^2 \times n_i$
1	6	6	6
2	7	14	28
3	4	12	36
4	2	8	32
5	1	5	25
Total	20	45	127

$$s^2 = \frac{\sum_1^5 x_i^2\, n_i - \frac{1}{20}(\sum_1^5 x_i\, n_i)^2}{20-1}$$

$$s^2 = \frac{127 - \frac{1}{20}45^2}{19} = 1.355$$

$$s = \sqrt{1.355} = 1.16$$

b.

En el caso de datos agrupados en intervalos de clase

c_i	n_i	$c_i \times n_i$	$c_i^2 \times n_i$
7.755	7	54.285	420.980
8.455	9	76.095	643.383
9.155	2	18.310	167.628
9.855	2	19.710	194.242
10.555	1	10.555	111.408
Total	21	178.955	1537.641

$$s^2 = \frac{\sum_1^5 c_i^2\, n_i - \frac{1}{20}(\sum_1^5 c_i\, n_i)^2}{21-1}$$

$$s^2 = \frac{1537.641 - \frac{1}{21}178.955^2}{20} = 0.632$$

$$s = \sqrt{0.632} = 0.795$$

(sin agrupar en intervalos se obtiene $s = 0.900$)

c.

En el ejemplo de las medidas con el péndulo simple, ya vimos que para el cálculo de la media aritmética efectuábamos un cambio de variable $y = a + b\,x = -980 + 100\,x$.

x_i	y_i
9.77	−3
9.78	−2
9.80	0
9.81	+1
9.83	+3
10.25	+45

$$s_x^2 = \frac{\sum_1^6 (x_i - \overline{x})^2}{N-1} \quad ; \quad s_y^2 = \frac{\sum_1^6 (y_i - \overline{y})^2}{N-1}$$

$$s_y^2 = \frac{\sum_1^6 (y_i - 7.33)^2}{5} = 345.07$$

$$\Rightarrow s_y = \sqrt{345.07} = 18.58$$

$$s_x = \frac{s_y}{b} = \frac{18.58}{100} = 0.186 \text{ m/s}^2.$$

Nótese que es mucho mayor que la desviación media $D_{\overline{x}} = 0.125$. La desviación típica es poco robusta y fuertemente dependiente de los valores extremos.

1.- Ahora hazlo tú.

I. Se tiene la muestra con los datos 700, 701, 701 y 702. Así como su = 701, su desviación será

II. Calcula la media, la varianza y la desviación típica tras encuestar a 25 familias sobre el número de hijos que tenían, se obtuvieron los siguientes datos

N° de hijos(X_i)	**0**	**1**	**2**	**3**	**4**	
N° de familias(n_i)	**5**	**6**	**8**	**4**	**2**	**25**

III. Del siguiente ejercicio calcular la varianza y la desviación típica

X	Intervalo	f.absoluta	f.acumulada	f.relativa	f.r.acumulada	f.x	x^2	f. x^2
52	50-54	7	7	0,078	0,078	364	2704	18928
56	54-58	10	17	0,111	0,189	560	3136	31360
60	58-62	16	33	0,178	0,367	960	3600	57600
64	62-66	20	53	0,222	0,589	1280	4096	81920
68	66-70	18	71	0,2	0,789	1224	4624	83232
72	70-74	11	82	0,122	0,911	792	5184	57024
76	74-78	8	90	0,089	1	608	5776	46208
448		90		1		5788		376272

IV. Halla la media y la desviación típica correspondientes a la siguiente distribución de edades:

Intervalo	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
Frecuencia	3	9	12	9	15	2

¿Qué porcentaje tienen menos de 15 años?

2.- Revisa los resultados obtenidos

I. 0,707…

II. La Varianza es: $s^2 = 4'24 - (1'68)^2 = 1'4176$.
Y la Desviación Típica s = 1'85.

III.

Varianza:

$$S^2 = [\Sigma f \cdot x^2 - [(\Sigma f \cdot x)^2 / N]] / (N - 1)$$

$$S^2 = [376272 - [(5788)^2 / 90]] / (90 - 1)$$

$S^2 = 45,402$.

Desviación típica:

(Raiz cuadrada de la varianza.)

S = 6,74

IV.

Hallamos la marca de clase, x_i, de cada intervalo y confeccionamos la tabla:

Intervalo	x_i	f_i	$x_i f_i$	$f_i x_i^2$
0 - 5	2,5	3	7,5	18,75
5 - 10	7,5	9	67,5	506,25
10 - 15	12,5	12	150	1875
15 - 20	17,5	9	157,5	2756,25
20 - 25	22,5	15	337,5	7593,75
25 - 30	27,5	2	55	1512,5
		50	775	14262,5

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{775}{50} = 15{,}5$$

$$\sigma = \sqrt{\frac{\Sigma f_i x_i^2}{n} - \bar{x}^2} = \sqrt{\frac{14262{,}5}{50} - 15{,}5^2} = \sqrt{45} = 6{,}71$$

La edad media del grupo es de 15,5 años, con una desviación típica de 6,71 años.

De 50 personas, $3 + 9 + 12 = 24$ tienen menos de 15 años. Por tanto:

$$\frac{24 \cdot 100}{50} = 48$$

Luego el 48% tienen menos de 15 años.

3.- Finalmente responde esta autoevaluación marcando la opción que corresponda luego de haber revisado tus respuestas.

Indicador	Sí	No
¿Realice el procedimiento algebraico para calcular la varianza?		
¿El valor obtenido es correcto?		
¿Realice el procedimiento algebraico para calcular la desviación típica?		