Desviación típica de una variable estadística

Considero una muestra de N individuos de una población. Se desea estudiar una característica de los individuos de la población que está definida por una variable estadística que toma N valores x_i.

Por ejemplo, considero una población formada por los 25 alumnos de una clase y la variable estadística que estudia las notas de un examen de Matemáticas.

x_i	1	2	3	4	5	6	7	8	9
F_i	2	2	4	5	5	3	2	1	1

$N = 25$ número de elementos de la muestra

x_i: Notas del examen de Matemáticas

F_i: frecuencias absolutas (número de alumnos que tienen la misma nota x_i)

Se define la media:

$$\bar{x} = \frac{\sum x_i \cdot F_i}{N}$$

En nuestro caso la nota media del examen es:

$$\bar{x} = \frac{1 \cdot 2 + 2 \cdot 2 + 3 \cdot 4 + 4 \cdot 5 + 5 \cdot 5 + 6 \cdot 3 + 7 \cdot 2 + 8 \cdot 1 + 9 \cdot 1}{25}$$

$$\bar{x} = \frac{2 + 4 + 12 + 20 + 25 + 18 + 14 + 8 + 9}{25} = \frac{112}{25} = 4{,}48$$

Se define la varianza:

$$V = \frac{\sum (x_i - \bar{x})^2 \cdot F_i}{N}$$

Y la desviación típica: $\sigma = \sqrt{V}$

La desviación típica mide la dispersión de los valores de la variable estadística con respecto a la media.

En general, la mayoría de los datos se encuentran en el intervalo $(\bar{x} - \sigma\ , \bar{x} + \sigma\ )$

En nuestro caso, la varianza y desviación típica se calculan con la tabla siguiente:

x_i	F_i	$\bar{x} - x_i$	$(\bar{x} - x_i)^2$	$(\bar{x} - x_i)^2 \cdot F_i$
1	2	3,48	12,11	24,22
2	2	2,48	6,15	12,3
3	4	1,48	2,19	8,76
4	5	0,48	0,23	1,15
5	5	-0,52	0,27	1,35
6	3	-1,52	2,31	6,93
7	2	-2,52	6,35	12,7
8	1	-3,52	12,39	12,39
9	1	-4,52	20,43	20,43
			Total:	100,23

La varianza es: $V = \frac{100,23}{25} = 4,0092$

y la desviación típica es $\sigma = \sqrt{4,0092} = 2,002$

El intervalo $(\bar{x} - \sigma\ , \bar{x} + \sigma\ ) = (4,48 - 2,002\ , 4,48 + 0,002) = (2,478\ , 6,482)$ contiene 17 notas del examen que representa el $\frac{17}{25} = 0,68 = 68\%$ de los alumnos de la clase.

Podemos afirmar que el 68% de los alumnos de la clase han obtenido notas superiores a 2 y inferiores a 7.

Regresión lineal

Consideremos dos variables estadísticas y_i y x_i definidas a partir de una muestra de N individuos de la población.

Por ejemplo, las notas de dos exámenes de Matemáticas de una clase de 20 alumnos.

x_i	4	3	4	6	3	5	6	1	5	3	6	4	4	6	3	5	2	4	5	6
y_i	5	2	6	7	4	4	8	2	5	2	6	3	5	7	4	5	3	3	4	7

Queremos saber, si es factible aproximar los resultados de la tabla por una función lineal

$$y = mx + b$$

De manera que se verifique:

$$y_i \approx mx_i + b$$

Esta función lineal se denomina recta de regresión.

Se define la covarianza σ_{xy} de las dos variables estadísticas:

$$\sigma_{xy} = \frac{\sum(x_i - \vec{x}) \cdot (y_i - \vec{y})}{N}$$

En nuestro caso,

$$\tilde{x} = \frac{1 \cdot 1 + 2 \cdot 1 + 3 \cdot 4 + 4 \cdot 5 + 5 \cdot 4 + 6 \cdot 5}{20} = 4{,}25$$

$$\tilde{y} = \frac{2 \cdot 3 + 3 \cdot 3 + 4 \cdot 4 + 5 \cdot 4 + 6 \cdot 2 + 7 \cdot 3 + 8 \cdot 1}{20} = 4{,}6$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
4	5	-0,25	0,4	-0,1
3	2	-1,25	-2,6	3,25
4	6	-0,25	1,4	-0,35
6	7	1,75	2,4	4,2
3	4	-1,25	-0,6	0.75
5	4	0,75	-0,6	-0,45
6	8	1,75	3,4	5,95
1	2	-3,25	-2,6	8.45
5	5	0,75	0,4	0,3
3	2	-1,25	-2,6	3.25
6	6	1,75	1,4	2,45
4	3	-0,25	-1,6	0,4
4	5	-0,25	0,4	-0,1
6	7	1,75	2,4	4,2
3	4	-1,25	0,4	-0,5
5	5	0,75	0,4	0,3
2	3	-2,25	-1,6	3,6
4	3	-0,25	-1,6	0,4
5	4	0,75	-0,6	-0,45
6	7	1,75	2,4	4,2
		Total:		39,75

Por tanto, la covarianza es $\sigma_{xy} = \frac{39,75}{20} = 1,9875$

Se define el coeficiente de correlación

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Se verifica que $-1 \leq \rho \leq 1$

Si $\rho \approx 1$ la recta de regresión aproxima bien los datos de las dos variables y la tendencia es creciente.

Si $\rho \approx -1$ la recta de regresión aproxima bien los datos de las dos variables y la tendencia es decreciente.

Si $\rho \approx 0$ los datos de las dos variables no están relacionados linealmente.

En nuestro caso:

x_i	F_i	$\bar{x} - x_i$	$(\bar{x} - x_i)^2$	$(\bar{x} - x_i)^2 \cdot F_i$
1	1	3,35	11,2225	11,2225
2	1	2,25	5,0625	5,0625
3	4	1,25	1,5625	6,25
4	5	0,25	0,0625	0,3125
5	4	-0,75	0,5625	2,25
6	5	-1,75	3,0625	15,3125
			Total:	40,41

$$\sigma_x = \sqrt{\frac{40{,}41}{20}} = 1{,}42$$

y_i	F_i	$\bar{\bar{y}} - y_i$	$(y - y_i)^2$	$(\bar{y} - y_i)^2 \cdot F_i$
2	3	2,6	6,76	20,28
3	3	1,6	2,56	7,68
4	4	0,6	0,36	1,44
5	4	-0,4	0,16	0,64
6	2	-1,4	1,96	3,92
7	3	-2,4	5,76	17,28
8	1	-3,4	11,56	11,56
			Total:	62,8

$$\sigma_y = \sqrt{\frac{62{,}8}{20}} = 1{,}77$$

El coeficiente de correlación es

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{1{,}9875}{1{,}42 \cdot 1{,}77} = 0{,}79$$

Se puede afirmar que hay una buena aproximación lineal con tendencia creciente de las notas de los dos exámenes.

La pendiente y la ordenada en el origen de la recta de regresión

$$y = mx + b$$

se calculan aplicando las fórmulas siguientes:

$$m = \frac{\sigma_{xy}}{{\sigma_x}^2} \qquad b = \vec{y} - \frac{\sigma_{xy}}{{\sigma_x}^2} \cdot \bar{x}$$

En nuestro caso

$$m = \frac{1{,}9875}{1{,}42^2} = 0{,}986 \qquad b = 4{,}6 - \frac{1{,}9875}{1{,}42^2} \cdot 4{,}25 = 0{,}41$$

La recta de regresión es

$$y = 0{,}986\, x + 0{,}41$$

En la representación gráfica de la nube de puntos se observa que la recta de regresión se ajusta a los datos