13.4 Linear Correlation

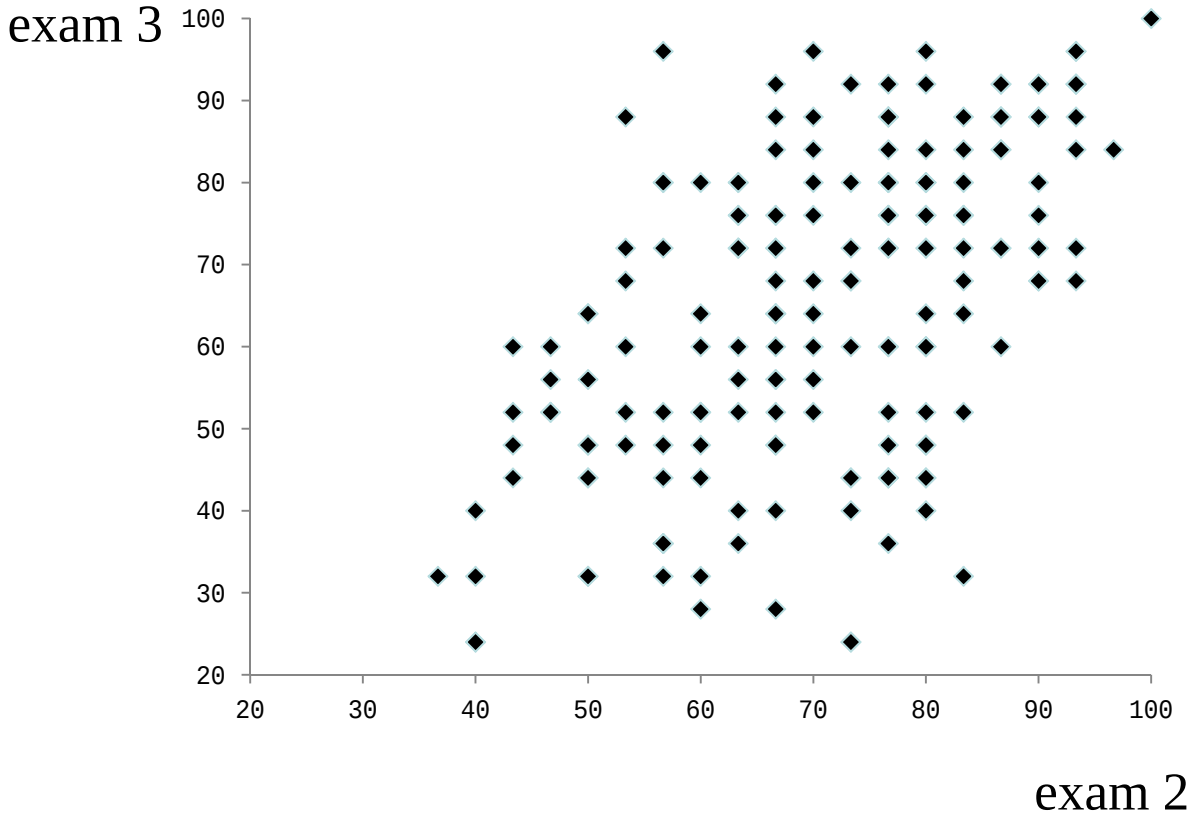
Definition. **Linear correlation coefficient** between two variables x and y is computed according to the formula:

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}}.$$

Properties of r

- r is a number such that $-1 \leq r \leq 1$.
- r is a measure of **direction** and **strength** of the **linear** relationship between x and y.
- $r > 0$ for positively associated variables, and $r < 0$ for negatively associated variables.
- The closer r is to -1 or 1, the stronger the linear relationship between x and y is. If $r = -1$ or $r = 1$, the relation is a perfect straight line.
- If $r = 0$, there is no linear relationship between x and y.

Example. Can we guess the value of r in this example?

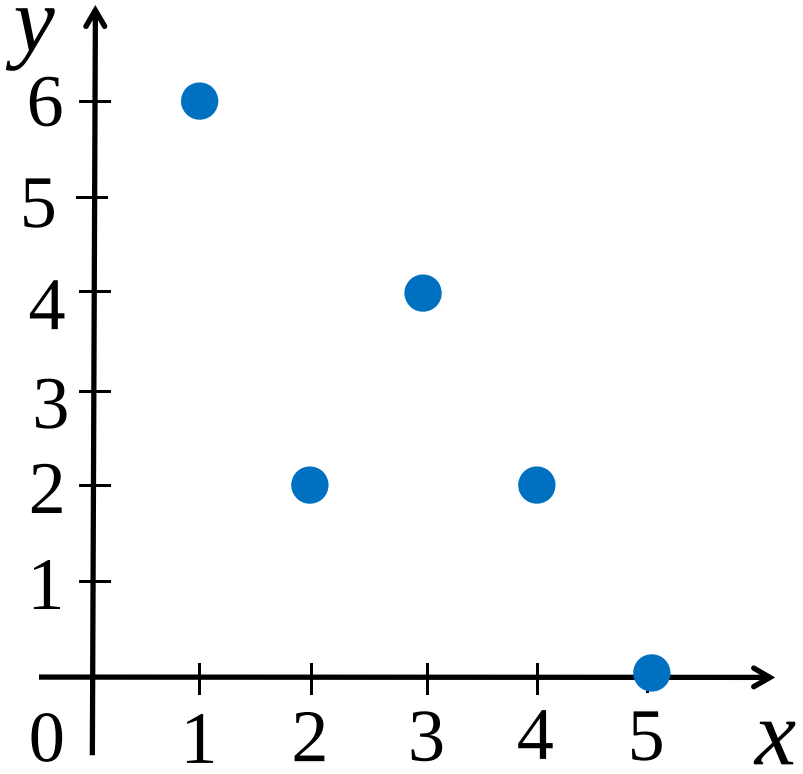


Answer. $r = 0.49$

Terminology.

correlation	linear relation	correlation	linear relation
$r = 0$	doesn't exist		
$-0.2 \le r < 0$	very weak negative	$0 < r \le 0.2$	very weak positive
$-0.5 \le r < -0.2$	weak negative	$0.2 < r \le 0.5$	weak positive
$-0.7 \le r < -0.5$	reasonably strong negative	$0.5 < r \le 0.7$	reasonably strong positive
$-0.9 \le r < -0.7$	strong negative	$0.7 < r \le 0.9$	strong positive
$-1 < r < -0.9$	very strong negative	$0.9 < r < 1$	very strong positive
$r = -1$	perfect negative	$r = 1$	perfect positive

Example. Suppose points (1,6), (2,2), (3,4), (4,2), and (5,0) are observed. Here is the scatter diagram. Find the linear correlation coefficient.



Solution. We compute r as follows:

$$\sum xy = (1)(6)+(2)(2)+(3)(4)+(4)(2)+(5)(0) = 30,$$

$$\sum x = 1+2+3+4+5 = 15,\ n = 5,\ \bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3,$$

$$\sum y = 6+2+4+2+0 = 14,\ \bar{y} = \frac{\sum y}{n} = \frac{14}{5} = 2.8,$$

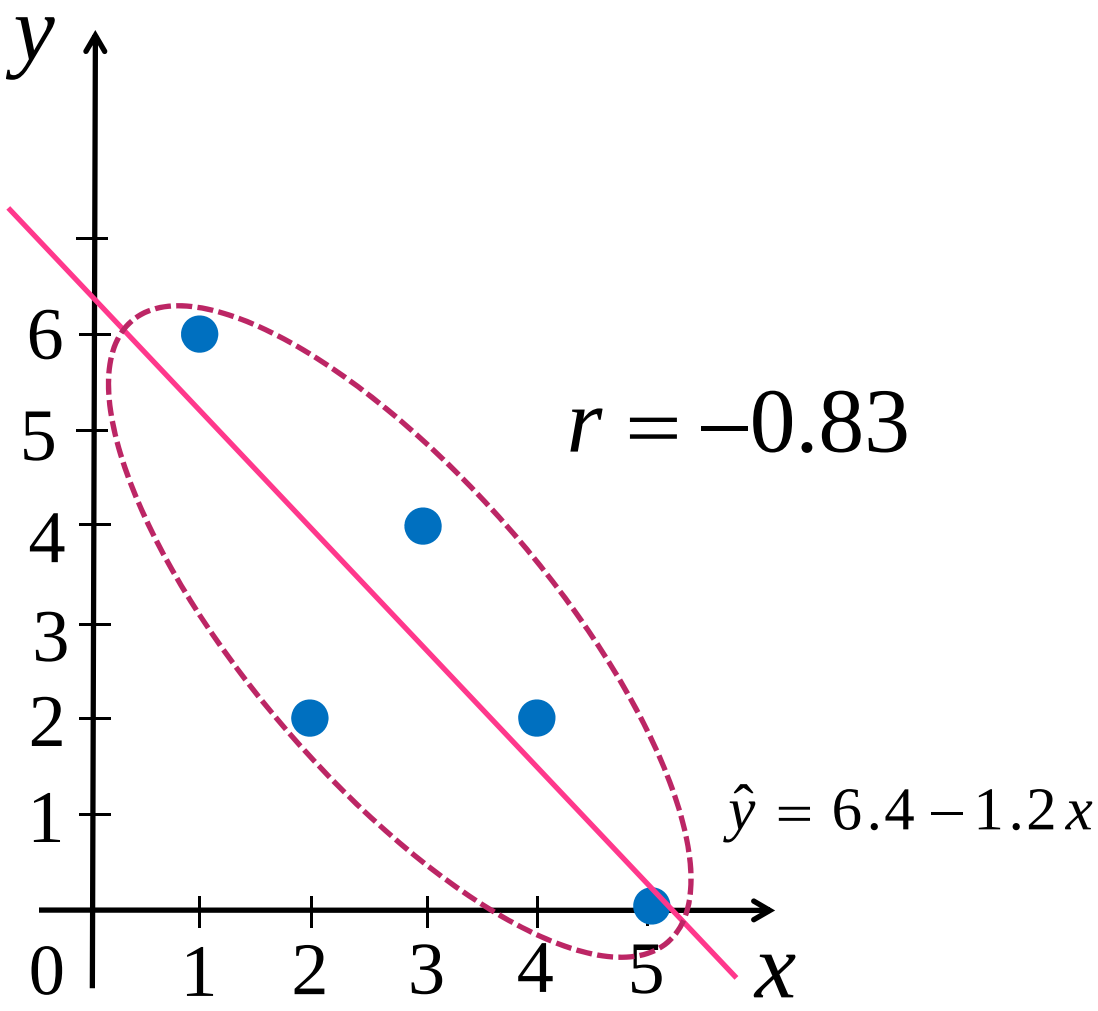
$$\sum x^2 = 1^2+2^2+3^2+4^2+5^2 = 1+4+9+16+25 = 55,$$

$$\sum y^2 = 6^2+2^2+4^2+2^2+0^2 = 36+4+16+4+0 = 60,$$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}} = \frac{30-(5)(3)(2.8)}{\sqrt{55-(5)(3)^2}\sqrt{60-(5)(2.8)^2}}$$

$$= \frac{-12}{\sqrt{10}\sqrt{20.8}} = -0.83.$$

The interpretation of $r = -0.83$ is that the **linear** relationship between x and y is **negative** and **strong** (the points on a scatter diagram lie close to the fitted regression line).


y
6
5
4
3
2
1
0
1
2
3
4
5
x
r = –0.83
ŷ = 6.4 – 1.2x