Variance, Covariance, and Correlation

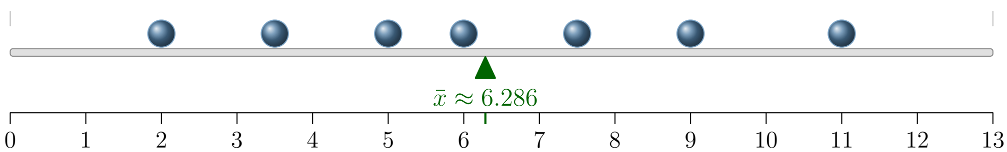
Mr. Merrick · September 29, 2025

1) Dataset and Means

Label	A	B	C	D	E	F	G	Totals
x_i	2.0	3.5	5.0	6.0	7.5	9.0	11.0	$\sum x_i = 44.0$
y_i	0.2	3.6	0.4	3.2	1.0	3.8	1.3	$\sum y_i = 13.5$

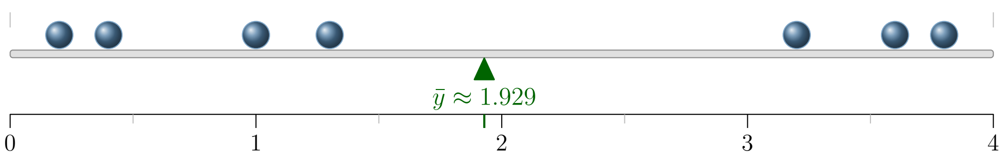
Think of each value as a small *weight* sitting on a beam. Without calculating, *eyeball* where the beam would balance and mark your guess on the ruler line below, and draw in a fulcrum.

Along the x-axis:



Mark the fulcrum at the balancing point ($\bar{x}$) under the beam. Calculate $\bar{x}$.

Along the y-axis:



Mark the fulcrum at the balancing point ($\bar{y}$) under the beam. Calculate $\bar{y}$.

Quick practice (Means)

1. On the balance beam, do spheres closer to the balance point or farther from it have a greater effect on where it balances? Why?

   Spheres farther from the balance point have a greater effect. Torque is weight $\times$ lever arm. With equal weights, the contribution to shifting the balance is proportional to the distance $|x_i - \bar{x}|$.

2. If every y_i is increased by the same constant a, how does the balance point on the y-beam move?

   It shifts up by a: if $y_i' = y_i + a$, then $\bar{y}' = \bar{y} + a$.

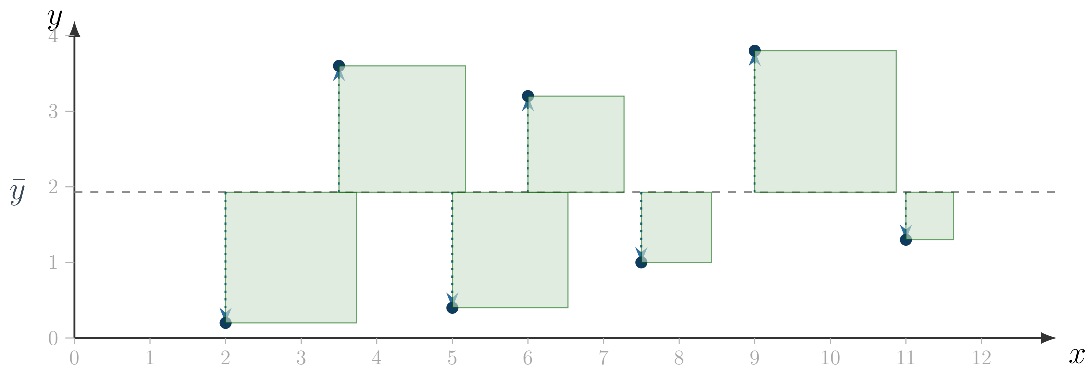
3. If all x-values are multiplied by a factor a (scaled), what happens to the balance point on the x-beam?

   It scales by the same factor: if $x_i' = a x_i$, then $\bar{x}' = a\,\bar{x}$. For $a > 0$ it stretches/compresses; for $a < 0$ it also reflects across 0.

We will use these same seven points in every section.

2) Variance of y (sample): average of squared deviations from mean

The horizontal dashed line is at $\bar{y} = 1.929$. Each dotted arrow has length $|y_i - \bar{y}|$. For every point, draw a **square** using that arrow as one side. Area $= (y_i - \bar{y})^2$. Your squares will overlap.



Variance in y (sample): $s_y^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$

Point	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
A	0.2	-1.729	2.988
B	3.6	1.671	2.794
C	0.4	-1.529	2.337
D	3.2	1.271	1.617
E	1.0	-0.929	0.862
F	3.8	1.871	3.502
G	1.3	-0.629	0.395
	$\sum y_i = 13.5$		14.494

Practice (Variance in y)

1. Which point lies farthest from the mean line (largest vertical deviation)? Which is closest? Explain using the diagram.

   Farthest: point F ($y = 3.8$) with $|y_i - \bar{y}| \approx 1.871$. Closest: point G ($y = 1.3$) with $|y_i - \bar{y}| \approx 0.629$. On the plot, F has the longest vertical dotted arrow from the mean line, and G has the shortest.

2. If every y_i were shifted upward by $+2$, would the variance s_y^2 change? Explain geometrically.

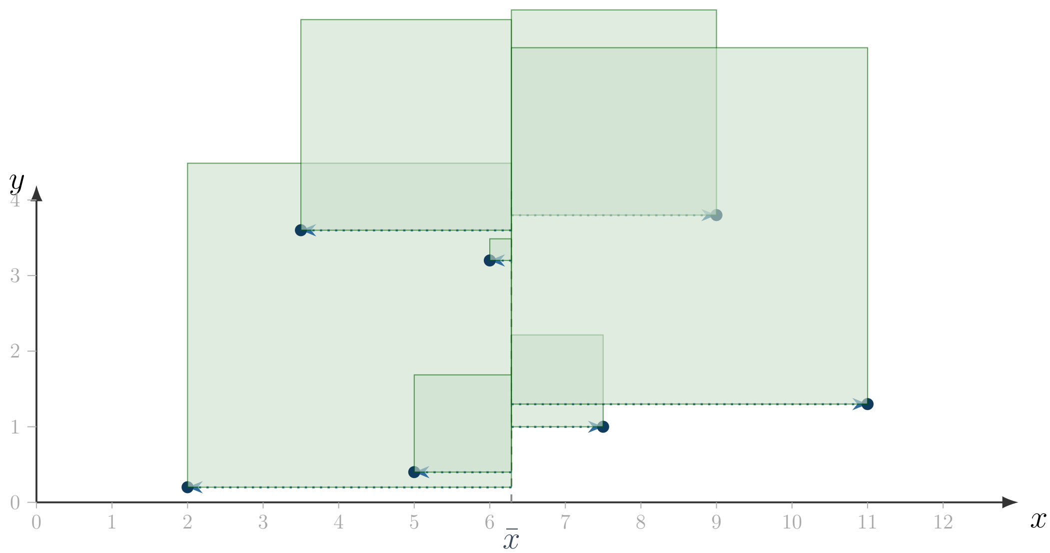
   No change. A uniform shift replaces each deviation by $y_i' - \bar{y}' = (y_i + 2) - (\bar{y} + 2) = y_i - \bar{y}$, so the squared deviations and their average stay the same. Geometrically, all arrows translate without changing lengths.

3. Compute the total sum of squares in y, $\mathrm{SST}_y = \sum (y_i - \bar{y})^2$. What proportion of this sum comes from points above the mean $\bar{y}$?

   $\mathrm{SST}_y \approx 14.494$. For the points above the mean (B, D, F): $2.794 + 1.617 + 3.502 = 7.913$. Proportion $\approx 7.913/14.494 \approx 0.546$ (about 54.6%).

3) Variance of x (sample): average of squared deviations from mean

The vertical dashed line is at $\bar{x} = 6.286$. Each dotted *horizontal* arrow has length $|x_i - \bar{x}|$. Draw squares using that arrow as one side. Area $= (x_i - \bar{x})^2$. Your squares will overlap.



Variance in x (sample): $s_x^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Point	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
A	2.0	-4.286	18.367
B	3.5	-2.786	7.760
C	5.0	-1.286	1.653
D	6.0	-0.286	0.082
E	7.5	1.214	1.474
F	9.0	2.714	7.367
G	11.0	4.714	22.224
	$\sum x_i = 44.0$		58.929

Practice (Variance in x)

1. Which points contribute most strongly to s_x^2? How can you tell just by looking at the diagram?

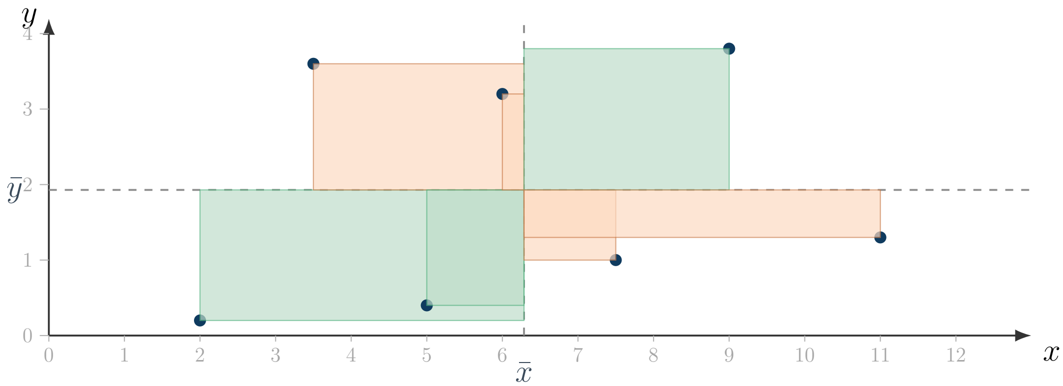
   Points farthest from $\bar{x}$ contribute most because each term is $(x_i - \bar{x})^2$. Here, G ($x = 11.0$) with $|x_i - \bar{x}| \approx 4.714$ and A ($x = 2.0$) with $|x_i - \bar{x}| \approx 4.286$ contribute the most; they have the longest horizontal dotted arrows.

2. If every x-value were rescaled by a factor k ($x_i' = kx_i$), how would the variance s_x^2 change?

   It scales by k^2: $s_{x'}^2 = \dfrac{1}{n-1} \sum (kx_i - k\bar{x})^2 = k^2 \dfrac{1}{n-1} \sum (x_i - \bar{x})^2 = k^2 s_x^2$. (For negative k, the sign flips but the square makes the factor k^2.)

4) Covariance (sample): average of signed rectangle areas

Draw a rectangle for each point with side lengths $|x_i - \bar{x}|$ and $|y_i - \bar{y}|$.
Quadrants I & III are positive; Quadrants II & IV are negative. Your rectangles will overlap.



Covariance (sample): $\mathrm{Cov}(X,Y) = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Point	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
A	2.0	0.2	-4.286	-1.729	7.408
B	3.5	3.6	-2.786	1.671	-4.656
C	5.0	0.4	-1.286	-1.529	1.965
D	6.0	3.2	-0.286	1.271	-0.363
E	7.5	1.0	1.214	-0.929	-1.128
F	9.0	3.8	2.714	1.871	5.080
G	11.0	1.3	4.714	-0.629	-2.963
	$\sum x_i = 44.0$	$\sum y_i = 13.5$			5.343

Practice (Covariance)

1. If you swapped the roles of x and y, would the covariance change? Why or why not?

   No change. Covariance is symmetric: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ because $(x_i - \bar{x})(y_i - \bar{y})$ is the same product either way.

2. For a scatterplot with a strong positive linear trend, what do you expect the sign and size of the covariance to be? What about a strong negative trend?

   Positive trend $\Rightarrow$ covariance > 0 (points with $x_i > \bar{x}$ tend to have $y_i > \bar{y}$ and vice versa). Negative trend $\Rightarrow$ covariance < 0. The tighter and more spread-out the cloud along the line, the larger the magnitude $|\mathrm{Cov}(X,Y)|$.

3. If all y values were doubled, how would the covariance change? Explain your reasoning.

   It doubles: $\mathrm{Cov}(X,2Y) = \frac{1}{n-1}\sum(x_i - \bar{x})\big(2(y_i - \bar{y})\big) = 2\ \mathrm{Cov}(X,Y)$. In general, scaling one variable by c scales covariance by c.

5) Correlation

After computing the sample variances and the sample covariance above, compute the (sample) correlation:

$$r = \frac{\mathrm{Cov}(X,Y)}{s_x\, s_y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y} \quad \text{where} \quad s_x = \sqrt{s_x^2}, \quad s_y = \sqrt{s_y^2}.$$

Summary table (from your work above):

	s_x^2	s_y^2	$\mathrm{Cov}(X,Y)$	$r = \frac{\mathrm{Cov}(X,Y)}{s_x s_y}$
Values	9.821	2.416	0.890	0.183

Practice (Correlation)

1. If x_i is measured in centimeters and y_i in grams, why might correlation (r) be easier to interpret than covariance?

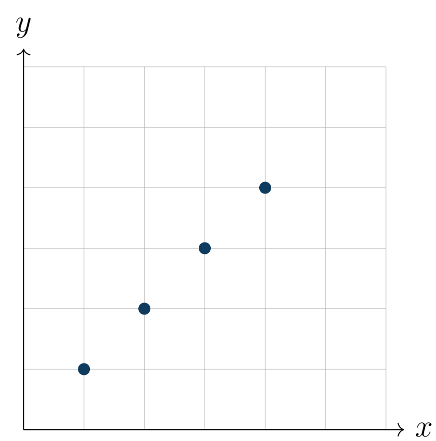
   Because correlation is unitless and always between -1 and 1. Covariance depends on the units (cm·g here) and can be hard to interpret in absolute terms. Correlation standardizes by s_x and s_y, making it scale-free.

2. Two datasets can have the same correlation r but look very different when graphed.

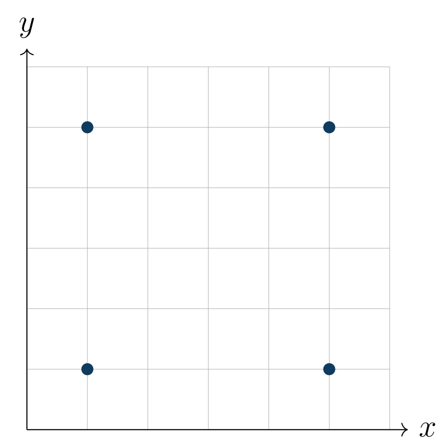
   For example, one dataset could be tightly clustered around a line, while another could be more spread out but still linear. Or one could have two distinct subgroups aligned along the same slope. Both give the same r, but the shape and distribution differ.

3. Draw two scatterplots with 4 points each: one with correlation $r = 1$ (perfect positive linear relationship), and one with correlation $r = 0$ (no linear relationship).

   For $r = 1$, all four points lie exactly on an increasing straight line. For $r = 0$, the points are arranged so there is no linear trend (e.g. a square or cross shape).



$r = 1$



$r = 0$

4. If x is rescaled from centimeters to meters, how does the correlation r change (if at all)? Explain.

   It does not change. Correlation is invariant under positive rescaling of either variable: multiplying all x_i by a constant rescales both numerator and denominator equally, leaving r unchanged.