Variance, Covariance, and Correlation

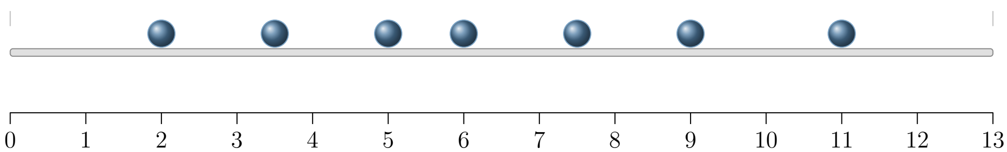
Mr. Merrick · September 29, 2025

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	$\sum y_i = 13.5$

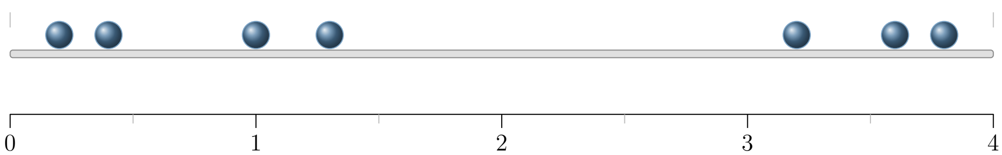
Think of each value as a small *weight* sitting on a beam. Without calculating, *eyeball* where the beam would balance and mark your guess on the ruler line below, and draw in a fulcrum.

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?

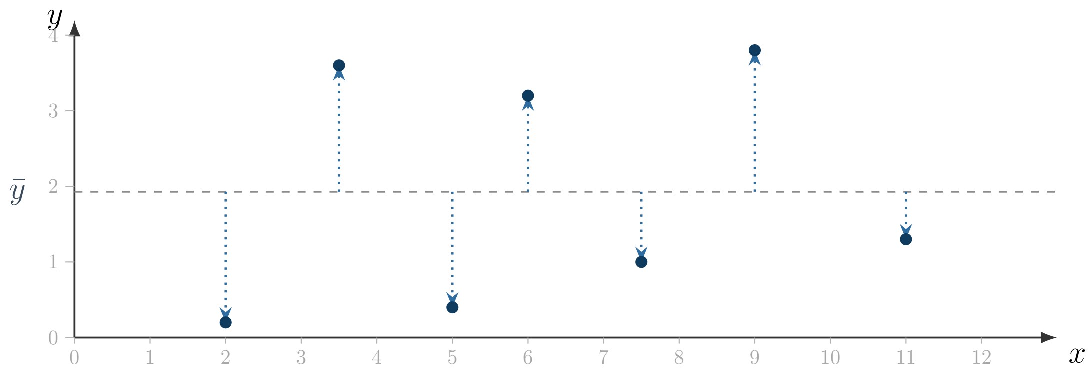
2. If every y_i is increased by the same constant a, how does the balance point on the y-beam move?

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

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		
B	3.6		
C	0.4		
D	3.2		
E	1.0		
F	3.8		
G	1.3		
	$\sum y_i = 13.5$		

Practice (Variance in y)

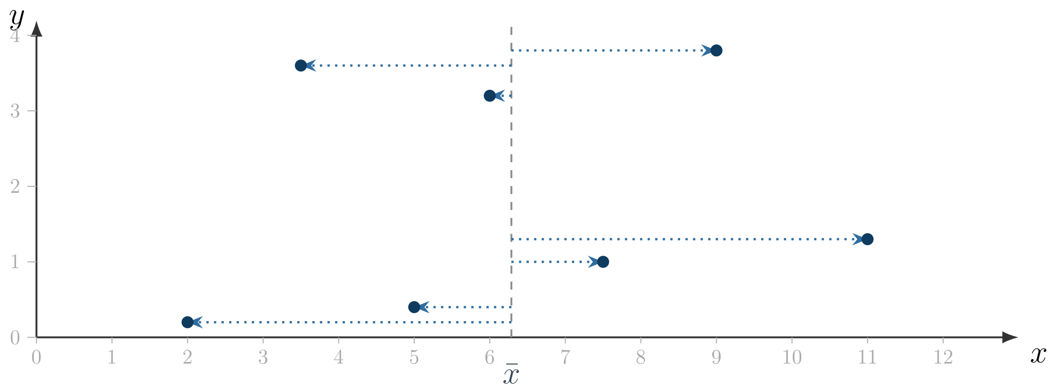
1. Which point lies farthest from the mean line (largest vertical deviation)? Which is closest? Explain using the diagram.

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

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

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		
B	3.5		
C	5.0		
D	6.0		
E	7.5		
F	9.0		
G	11.0		
	$\sum x_i = 44.0$		

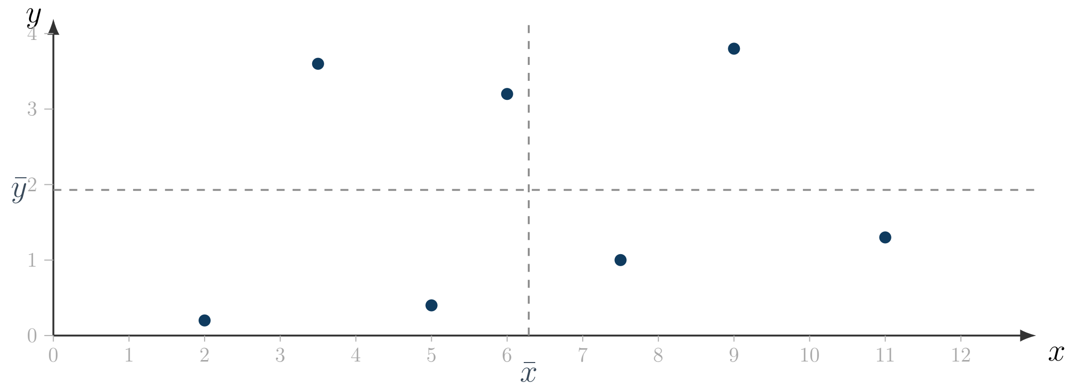
Practice (Variance in x)

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

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

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

Practice (Covariance)

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

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?

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

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):

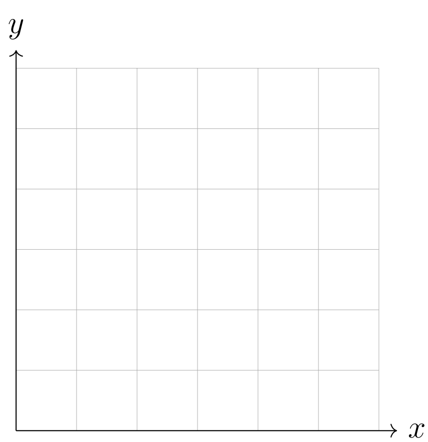
	s_x^2	s_y^2	$\mathrm{Cov}(X,Y)$	$r = \dfrac{\mathrm{Cov}(X,Y)}{s_x s_y}$
Values				

Practice (Correlation)

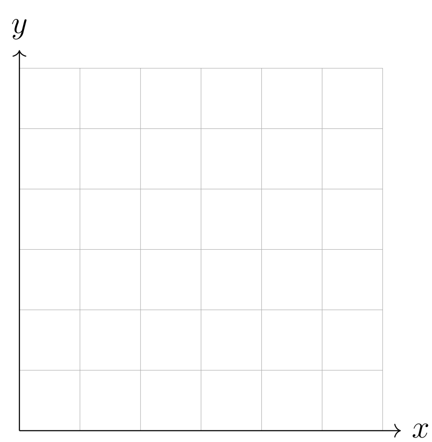
1. If x_i is measured in centimeters and y_i in grams, why might correlation (r) be easier to interpret than covariance?

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

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).



$r = 1$



$r = 0$

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