

Chapter 8: Descriptive Statistics

Bivariate Data

What changes in Chapter 8?

In Chapter 7, each observation gave us **one** number.

What changes in Chapter 8?

In Chapter 7, each observation gave us **one** number.

Now each observation gives us **two** numbers:

$$(x, y).$$

What changes in Chapter 8?

In Chapter 7, each observation gave us **one** number.
Now each observation gives us **two** numbers:

$$(x, y).$$

Examples:

- ▶ height and weight,
- ▶ exam 1 score and exam 2 score,
- ▶ distance traveled and airfare,
- ▶ automobile weight and fuel economy.

What changes in Chapter 8?

In Chapter 7, each observation gave us **one** number.
Now each observation gives us **two** numbers:

$$(x, y).$$

Examples:

- ▶ height and weight,
- ▶ exam 1 score and exam 2 score,
- ▶ distance traveled and airfare,
- ▶ automobile weight and fuel economy.

This is called **bivariate data**.

The main question

With bivariate numerical data, we are usually asking:

Is there a relationship between x and y ?

The main question

With bivariate numerical data, we are usually asking:

Is there a relationship between x and y ?

More specifically:

- ▶ Do larger x -values tend to go with larger y -values?
- ▶ Do larger x -values tend to go with smaller y -values?
- ▶ Is the relationship strong or weak?
- ▶ Is the relationship linear?
- ▶ Does the relationship suggest causation?

Chapter roadmap

By the end of this chapter, students should be able to:

- 1 construct and interpret scatter plots;
- 2 distinguish positive, negative, weak, and strong association;
- 3 understand covariance;
- 4 compute and interpret the correlation coefficient r ;
- 5 recognize that r measures **linear** association;
- 6 understand how shifts and rescaling affect correlation;
- 7 distinguish association from causation;
- 8 recognize hidden variables and influential points.

Skill 1: Construct a scatter plot

For every observation

$$(x_i, y_i),$$

plot one point in the coordinate plane.

Skill 1: Construct a scatter plot

For every observation

$$(x_i, y_i),$$

plot one point in the coordinate plane.

The first variable goes on the horizontal axis:

$$x.$$

Skill 1: Construct a scatter plot

For every observation

$$(x_i, y_i),$$

plot one point in the coordinate plane.

The first variable goes on the horizontal axis:

x .

The second variable goes on the vertical axis:

y .

Skill 1: Construct a scatter plot

For every observation

$$(x_i, y_i),$$

plot one point in the coordinate plane.

The first variable goes on the horizontal axis:

x .

The second variable goes on the vertical axis:

y .

Usually, the points are **not connected**.

Example 1A: Turn a table into points

Suppose we record hours studied and exam score.

Hours (x)	1	2	3	4	5
Score (y)	55	63	68	76	84

Example 1A: Turn a table into points

Suppose we record hours studied and exam score.

Hours (x)	1	2	3	4	5
Score (y)	55	63	68	76	84

The observations are:

$(1, 55)$, $(2, 63)$, $(3, 68)$, $(4, 76)$, $(5, 84)$.

Example 1A: Turn a table into points

Suppose we record hours studied and exam score.

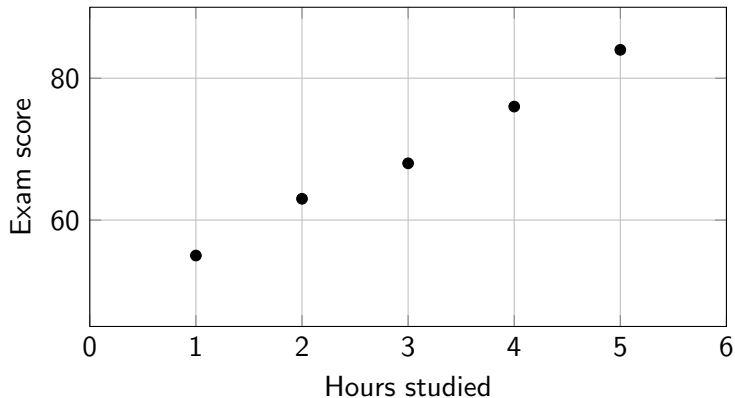
Hours (x)	1	2	3	4	5
Score (y)	55	63	68	76	84

The observations are:

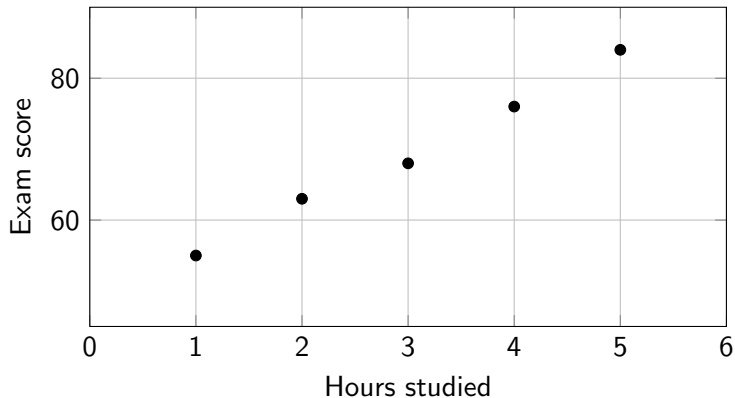
$(1, 55)$, $(2, 63)$, $(3, 68)$, $(4, 76)$, $(5, 84)$.

Each column gives **one point**.

Example 1A: Scatter plot



Example 1A: Scatter plot



The points clearly trend **upward from left to right**.

What do we look for in a scatter plot?

Use three questions:

- 1 **Direction:** positive or negative?

What do we look for in a scatter plot?

Use three questions:

- 1 **Direction:** positive or negative?
- 2 **Strength:** strong, moderate, or weak?

What do we look for in a scatter plot?

Use three questions:

- ① **Direction:** positive or negative?
- ② **Strength:** strong, moderate, or weak?
- ③ **Form:** approximately linear, or clearly curved/nonlinear?

What do we look for in a scatter plot?

Use three questions:

- ① **Direction:** positive or negative?
- ② **Strength:** strong, moderate, or weak?
- ③ **Form:** approximately linear, or clearly curved/nonlinear?

Also look for:

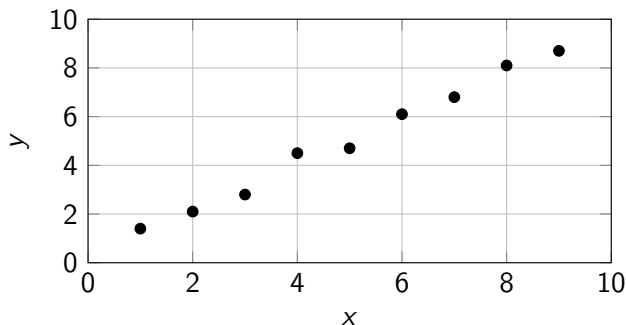
- ▶ unusual points,
- ▶ clusters or subgroups.

Positive association

A scatter plot has a **positive association** when larger x -values tend to occur with larger y -values.

Positive association

A scatter plot has a **positive association** when larger x -values tend to occur with larger y -values.

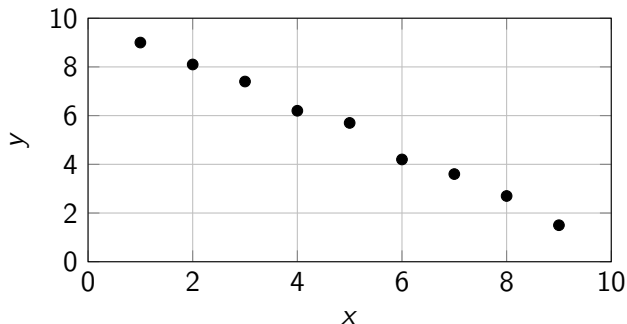


Negative association

A scatter plot has a **negative association** when larger x -values tend to occur with smaller y -values.

Negative association

A scatter plot has a **negative association** when larger x -values tend to occur with smaller y -values.



Example 1B: Identify the direction

Suppose the data are:

$(1, 9), (2, 7), (3, 8), (4, 5), (5, 4), (6, 2).$

Example 1B: Identify the direction

Suppose the data are:

$(1, 9), (2, 7), (3, 8), (4, 5), (5, 4), (6, 2).$

As x generally increases, y generally decreases.

Example 1B: Identify the direction

Suppose the data are:

$(1, 9), (2, 7), (3, 8), (4, 5), (5, 4), (6, 2).$

As x generally increases, y generally decreases.

Therefore the association is:

negative.

Example 1B: Identify the direction

Suppose the data are:

$(1, 9), (2, 7), (3, 8), (4, 5), (5, 4), (6, 2).$

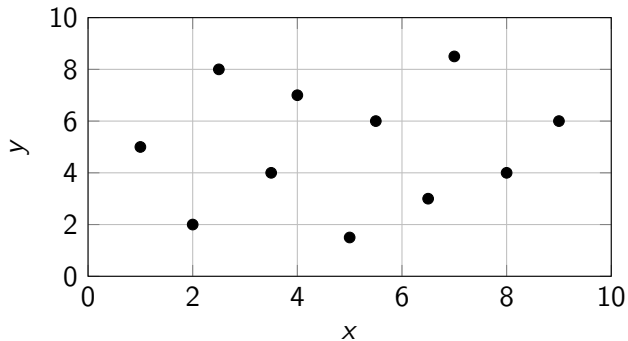
As x generally increases, y generally decreases.

Therefore the association is:

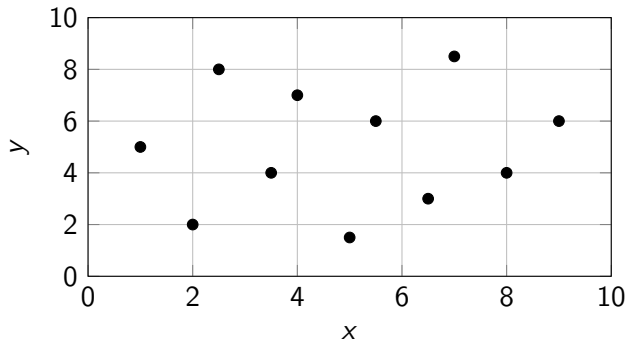
negative.

It is not necessary for *every* point to follow the pattern perfectly.

Example 1C: No clear linear trend



Example 1C: No clear linear trend



There is no obvious upward or downward **linear** pattern.

Skill 2: Why covariance?

A scatter plot gives a visual impression.

Skill 2: Why covariance?

A scatter plot gives a visual impression.

We also want a number that tells us whether x and y tend to move in the same direction.

Skill 2: Why covariance?

A scatter plot gives a visual impression.

We also want a number that tells us whether x and y tend to move in the same direction.

For each point (x_i, y_i) , compare it with the two means:

$$\bar{x}, \quad \bar{y}.$$

Skill 2: Why covariance?

A scatter plot gives a visual impression.

We also want a number that tells us whether x and y tend to move in the same direction.

For each point (x_i, y_i) , compare it with the two means:

$$\bar{x}, \quad \bar{y}.$$

Then examine:

$$(x_i - \bar{x})(y_i - \bar{y}).$$

The sign of one covariance contribution

Consider

$$(x_i - \bar{x})(y_i - \bar{y}).$$

The sign of one covariance contribution

Consider

$$(x_i - \bar{x})(y_i - \bar{y}).$$

If both x_i and y_i are above their means:

$$(+)(+) = +.$$

The sign of one covariance contribution

Consider

$$(x_i - \bar{x})(y_i - \bar{y}).$$

If both x_i and y_i are above their means:

$$(+)(+) = +.$$

If both are below their means:

$$(-)(-) = +.$$

The sign of one covariance contribution

Consider

$$(x_i - \bar{x})(y_i - \bar{y}).$$

If both x_i and y_i are above their means:

$$(+)(+) = +.$$

If both are below their means:

$$(-)(-) = +.$$

If one is above and the other is below:

$$(+)(-) = -.$$

The sign of one covariance contribution

Consider

$$(x_i - \bar{x})(y_i - \bar{y}).$$

If both x_i and y_i are above their means:

$$(+)(+) = +.$$

If both are below their means:

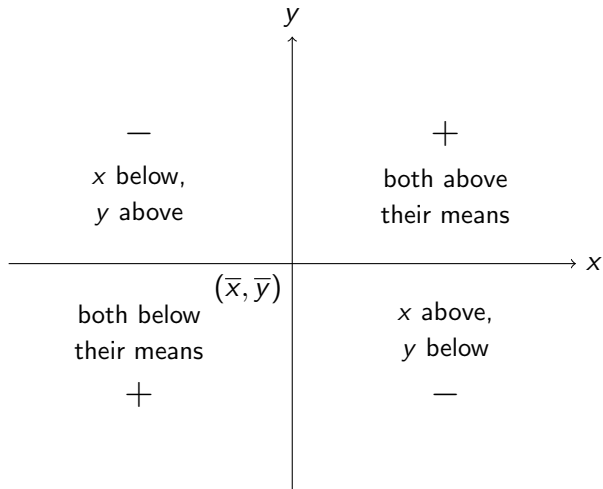
$$(-)(-) = +.$$

If one is above and the other is below:

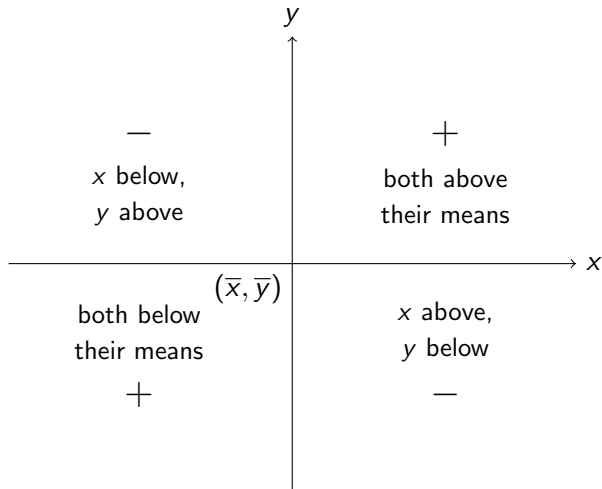
$$(+)(-) = -.$$

So positive products support a positive trend; negative products oppose it.

Four regions around $(\bar{x}, \bar{y})$



Four regions around $(\bar{x}, \bar{y})$



A positive linear trend tends to place many points in the two “+” regions.

Definition of covariance

For n data pairs,

$$c_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Definition of covariance

For n data pairs,

$$c_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Interpret the **sign**:

$c_{xy} > 0 \Rightarrow$ positive tendency,

$c_{xy} < 0 \Rightarrow$ negative tendency.

Definition of covariance

For n data pairs,

$$c_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Interpret the **sign**:

$c_{xy} > 0 \Rightarrow$ positive tendency,

$c_{xy} < 0 \Rightarrow$ negative tendency.

A covariance near 0 suggests little net **linear** tendency.

Example 2A: Covariance by hand

Use the small data set

$(1, 2)$, $(2, 4)$, $(3, 5)$.

Example 2A: Covariance by hand

Use the small data set

$$(1, 2), (2, 4), (3, 5).$$

First find the means:

$$\bar{x} = \frac{1 + 2 + 3}{3} = 2,$$

$$\bar{y} = \frac{2 + 4 + 5}{3} = \frac{11}{3}.$$

Example 2A: Build the products

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	-1	$-\frac{5}{3}$	$\frac{5}{3}$
2	4	0	$\frac{1}{3}$	0
3	5	1	$\frac{4}{3}$	$\frac{4}{3}$

Example 2A: Build the products

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	-1	$-\frac{5}{3}$	$\frac{5}{3}$
2	4	0	$\frac{1}{3}$	0
3	5	1	$\frac{4}{3}$	$\frac{4}{3}$

Add the products:

$$\frac{5}{3} + 0 + \frac{4}{3} = 3.$$

Example 2A: Build the products

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	-1	$-\frac{5}{3}$	$\frac{5}{3}$
2	4	0	$\frac{1}{3}$	0
3	5	1	$\frac{4}{3}$	$\frac{4}{3}$

Add the products:

$$\frac{5}{3} + 0 + \frac{4}{3} = 3.$$

Divide by $n - 1 = 2$:

$$c_{xy} = \frac{3}{2} = \boxed{1.5}.$$

Example 2A: Build the products

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	-1	$-\frac{5}{3}$	$\frac{5}{3}$
2	4	0	$\frac{1}{3}$	0
3	5	1	$\frac{4}{3}$	$\frac{4}{3}$

Add the products:

$$\frac{5}{3} + 0 + \frac{4}{3} = 3.$$

Divide by $n - 1 = 2$:

$$c_{xy} = \frac{3}{2} = \boxed{1.5}.$$

The covariance is positive, matching the upward trend.

Example 2B: Negative covariance

Data:

$(1, 8), (2, 6), (3, 4), (4, 2).$

Example 2B: Negative covariance

Data:

$(1, 8), (2, 6), (3, 4), (4, 2).$

Here, as x increases, y decreases perfectly.

Example 2B: Negative covariance

Data:

$(1, 8), (2, 6), (3, 4), (4, 2).$

Here, as x increases, y decreases perfectly.

So every point supports a negative relationship.

Example 2B: Negative covariance

Data:

$(1, 8), (2, 6), (3, 4), (4, 2).$

Here, as x increases, y decreases perfectly.

So every point supports a negative relationship.

Therefore:

$$c_{xy} < 0.$$

Example 2B: Negative covariance

Data:

$$(1, 8), (2, 6), (3, 4), (4, 2).$$

Here, as x increases, y decreases perfectly.

So every point supports a negative relationship.

Therefore:

$$c_{xy} < 0.$$

In fact, the covariance is

$$-\frac{10}{3}.$$

Why covariance is not enough

The **sign** of covariance is useful.

Why covariance is not enough

The **sign** of covariance is useful.

But the **size** of covariance depends on the units.

Why covariance is not enough

The **sign** of covariance is useful.

But the **size** of covariance depends on the units.

Suppose:

x = height in inches, y = weight in pounds.

If we convert height from inches to feet, the scatter plot has the same basic pattern.

Why covariance is not enough

The **sign** of covariance is useful.

But the **size** of covariance depends on the units.

Suppose:

x = height in inches, y = weight in pounds.

If we convert height from inches to feet, the scatter plot has the same basic pattern.

But the covariance changes numerically.

Why covariance is not enough

The **sign** of covariance is useful.

But the **size** of covariance depends on the units.

Suppose:

x = height in inches, y = weight in pounds.

If we convert height from inches to feet, the scatter plot has the same basic pattern.

But the covariance changes numerically.

So covariance is not a standardized measure of strength.

Skill 3: Correlation coefficient

The correlation coefficient standardizes covariance:

$$r = \frac{c_{xy}}{s_x s_y}$$

where:

s_x = standard deviation of the x -values,

s_y = standard deviation of the y -values.

Skill 3: Correlation coefficient

The correlation coefficient standardizes covariance:

$$r = \frac{c_{xy}}{s_x s_y}$$

where:

s_x = standard deviation of the x -values,

s_y = standard deviation of the y -values.

Unlike covariance, r has **no units**.

The most important fact about r

For every bivariate data set:

$$-1 \leq r \leq 1.$$

The most important fact about r

For every bivariate data set:

$$-1 \leq r \leq 1.$$

The sign gives the **direction**.

The most important fact about r

For every bivariate data set:

$$-1 \leq r \leq 1.$$

The sign gives the **direction**.

The distance from 0 gives the **strength of the linear relationship**.

$r \approx 1 \Rightarrow$ strong positive linear association

Interpreting r

$r \approx 1 \Rightarrow$ strong positive linear association

$r \approx -1 \Rightarrow$ strong negative linear association

Interpreting r

$r \approx 1 \Rightarrow$ strong positive linear association

$r \approx -1 \Rightarrow$ strong negative linear association

$r \approx 0 \Rightarrow$ weak or no linear association

Interpreting r

$r \approx 1 \Rightarrow$ strong positive linear association

$r \approx -1 \Rightarrow$ strong negative linear association

$r \approx 0 \Rightarrow$ weak or no linear association

The word **linear** matters.

A rough interpretation scale

There is no universal cutoff, but a useful classroom guide is:

$ r $	rough description
0 to 0.3	weak
0.3 to 0.7	moderate
0.7 to 1	strong

A rough interpretation scale

There is no universal cutoff, but a useful classroom guide is:

$ r $	rough description
0 to 0.3	weak
0.3 to 0.7	moderate
0.7 to 1	strong

Always inspect the scatter plot before trusting a label based only on r .

Example 3A: Compute r from summary statistics

Suppose:

$$c_{xy} = 12, \quad s_x = 3, \quad s_y = 5.$$

Example 3A: Compute r from summary statistics

Suppose:

$$c_{xy} = 12, \quad s_x = 3, \quad s_y = 5.$$

Use

$$r = \frac{c_{xy}}{s_x s_y}.$$

Example 3A: Compute r from summary statistics

Suppose:

$$c_{xy} = 12, \quad s_x = 3, \quad s_y = 5.$$

Use

$$r = \frac{c_{xy}}{s_x s_y}.$$

$$r = \frac{12}{(3)(5)} = \frac{12}{15} = \boxed{0.80}.$$

Example 3A: Compute r from summary statistics

Suppose:

$$c_{xy} = 12, \quad s_x = 3, \quad s_y = 5.$$

Use

$$r = \frac{c_{xy}}{s_x s_y}.$$

$$r = \frac{12}{(3)(5)} = \frac{12}{15} = \boxed{0.80}.$$

Interpretation:

strong positive linear association.

Example 3B: Negative correlation

Suppose:

$$c_{xy} = -18, \quad s_x = 4, \quad s_y = 6.$$

Example 3B: Negative correlation

Suppose:

$$c_{xy} = -18, \quad s_x = 4, \quad s_y = 6.$$

$$r = \frac{-18}{(4)(6)} = \frac{-18}{24}.$$

Example 3B: Negative correlation

Suppose:

$$c_{xy} = -18, \quad s_x = 4, \quad s_y = 6.$$

$$r = \frac{-18}{(4)(6)} = \frac{-18}{24}.$$

$$\boxed{r = -0.75}.$$

Example 3B: Negative correlation

Suppose:

$$c_{xy} = -18, \quad s_x = 4, \quad s_y = 6.$$

$$r = \frac{-18}{(4)(6)} = \frac{-18}{24}.$$

$$r = -0.75.$$

Interpretation:

strong negative linear association.

Example 3C: Perfect positive correlation

Data:

$(1, 2), (2, 4), (3, 6), (4, 8).$

Example 3C: Perfect positive correlation

Data:

$(1, 2), (2, 4), (3, 6), (4, 8).$

Every point lies exactly on the line

$$y = 2x.$$

Example 3C: Perfect positive correlation

Data:

$$(1, 2), (2, 4), (3, 6), (4, 8).$$

Every point lies exactly on the line

$$y = 2x.$$

Therefore:

$$\boxed{r = 1}.$$

Example 3C: Perfect positive correlation

Data:

$$(1, 2), (2, 4), (3, 6), (4, 8).$$

Every point lies exactly on the line

$$y = 2x.$$

Therefore:

$$\boxed{r = 1}.$$

Perfect positive correlation means all points lie exactly on a line with positive slope.

Example 3D: Perfect negative correlation

Data:

$(1, 8), (2, 6), (3, 4), (4, 2).$

Example 3D: Perfect negative correlation

Data:

$$(1, 8), (2, 6), (3, 4), (4, 2).$$

Every point lies exactly on

$$y = 10 - 2x.$$

Example 3D: Perfect negative correlation

Data:

$$(1, 8), (2, 6), (3, 4), (4, 2).$$

Every point lies exactly on

$$y = 10 - 2x.$$

Therefore:

$$\boxed{r = -1}.$$

Example 3D: Perfect negative correlation

Data:

$$(1, 8), (2, 6), (3, 4), (4, 2).$$

Every point lies exactly on

$$y = 10 - 2x.$$

Therefore:

$$\boxed{r = -1}.$$

Perfect negative correlation means all points lie exactly on a line with negative slope.

Skill 4: Match a scatter plot to r

Do this in order:

- 1 Is the trend upward or downward?

Skill 4: Match a scatter plot to r

Do this in order:

- 1 Is the trend upward or downward?
- 2 How tightly do the points follow a straight line?

Skill 4: Match a scatter plot to r

Do this in order:

- ① Is the trend upward or downward?
- ② How tightly do the points follow a straight line?
- ③ Is the relationship actually curved instead?

Skill 4: Match a scatter plot to r

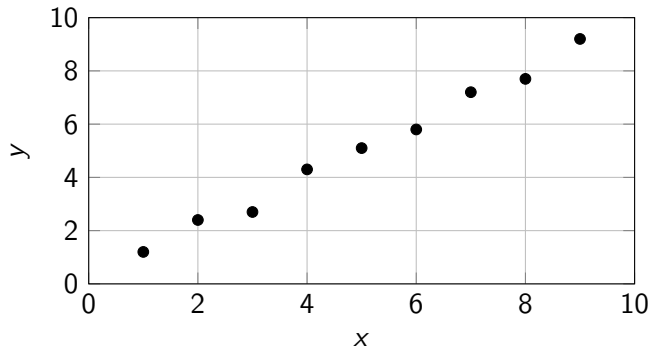
Do this in order:

- ① Is the trend upward or downward?
- ② How tightly do the points follow a straight line?
- ③ Is the relationship actually curved instead?

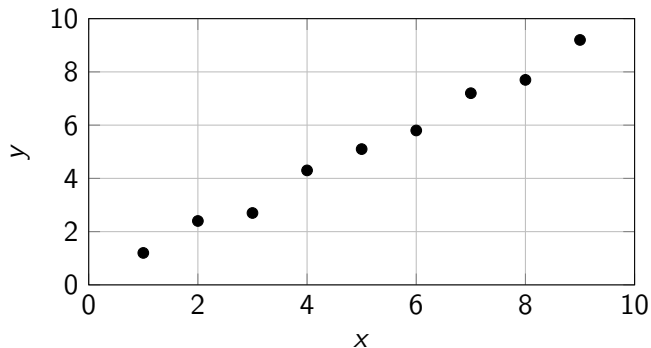
Direction tells you the **sign**.

Tightness around a line tells you the approximate **magnitude**.

Example 4A: Strong positive



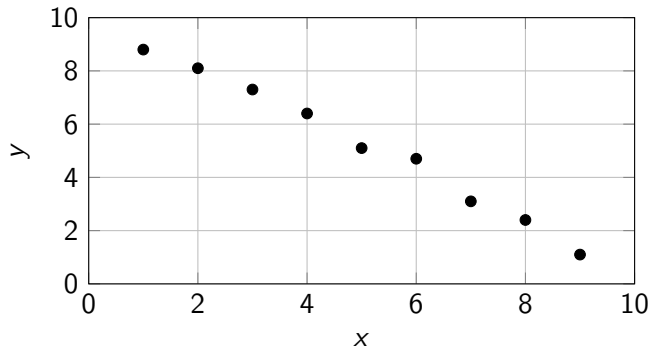
Example 4A: Strong positive



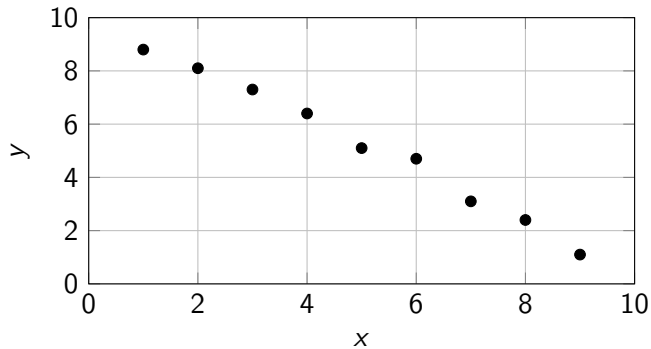
The trend is upward and points are close to a line.

r should be positive and close to 1.

Example 4B: Strong negative



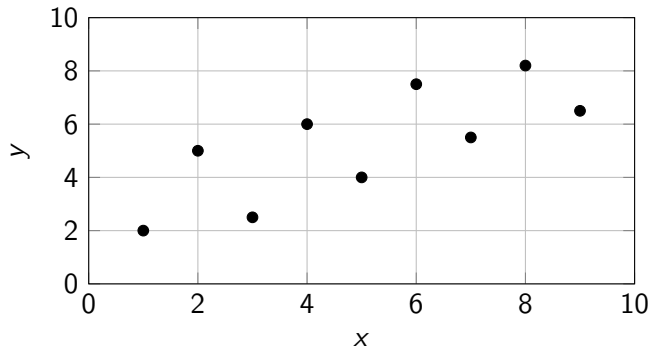
Example 4B: Strong negative



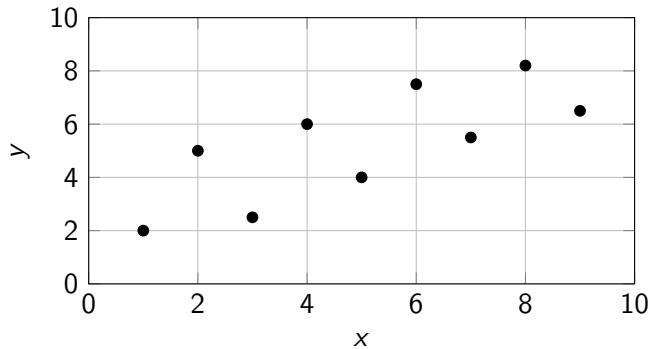
The trend is downward and points are close to a line.

r should be negative and close to -1 .

Example 4C: Moderate positive



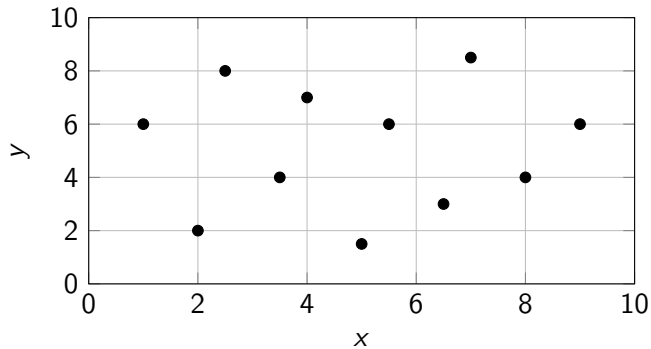
Example 4C: Moderate positive



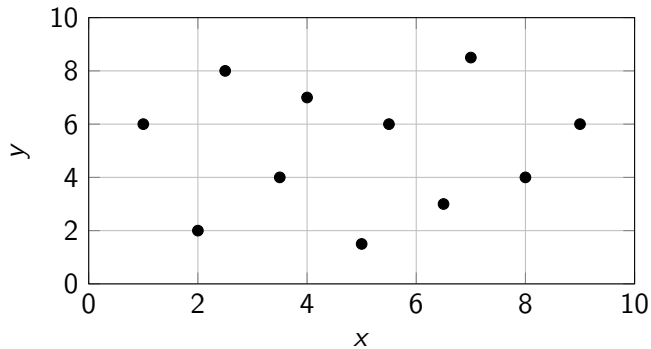
There is an upward pattern, but considerable scatter.

r is positive but not especially close to 1.

Example 4D: Weak linear association



Example 4D: Weak linear association



The points form a broad cloud.

r should be fairly close to 0.

Skill 5: A major warning

A correlation near 0 does **not** mean:

“there is no relationship.”

Skill 5: A major warning

A correlation near 0 does **not** mean:

“there is no relationship.”

It means:

there is little or no *linear* relationship.

Skill 5: A major warning

A correlation near 0 does **not** mean:

“there is no relationship.”

It means:

there is little or no *linear* relationship.

A strong curved relationship can still have

$$r \approx 0.$$

Example 5A: Strong relationship, but $r = 0$

Consider:

$$(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4).$$

Example 5A: Strong relationship, but $r = 0$

Consider:

$$(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4).$$

These points lie exactly on

$$y = x^2.$$

Example 5A: Strong relationship, but $r = 0$

Consider:

$$(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4).$$

These points lie exactly on

$$y = x^2.$$

The relationship is extremely clear.

Example 5A: Strong relationship, but $r = 0$

Consider:

$$(-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4).$$

These points lie exactly on

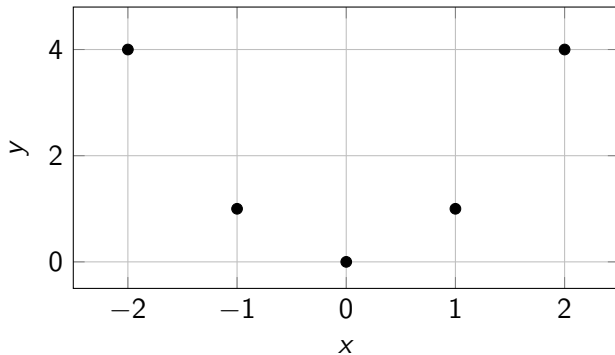
$$y = x^2.$$

The relationship is extremely clear.

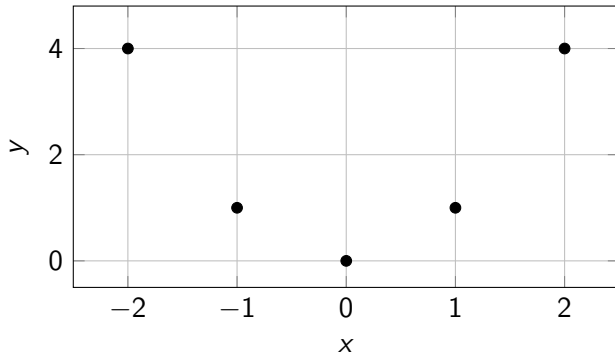
But:

$$\boxed{r = 0}.$$

Example 5A: See the curve

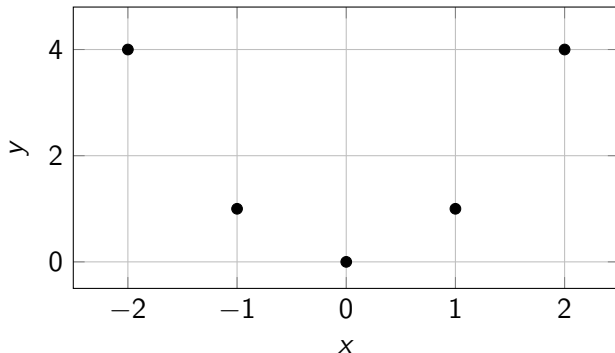


Example 5A: See the curve



The positive and negative linear tendencies cancel.

Example 5A: See the curve



The positive and negative linear tendencies cancel.
Therefore:

$$r = 0$$

even though the pattern is obviously not random.

Example 5B: Same r , very different pictures

Two data sets can have similar correlation coefficients but very different scatter plots.

Example 5B: Same r , very different pictures

Two data sets can have similar correlation coefficients but very different scatter plots.
Therefore:

Always look at the scatter plot.

Example 5B: Same r , very different pictures

Two data sets can have similar correlation coefficients but very different scatter plots.
Therefore:

Always look at the scatter plot.

The number r summarizes only one feature:

strength and direction of linear association.

Skill 6: Adding constants

Suppose we replace every x by

$$x + 100$$

and every y by

$$y - 20.$$

Skill 6: Adding constants

Suppose we replace every x by

$$x + 100$$

and every y by

$$y - 20.$$

The scatter plot shifts horizontally and vertically.

Skill 6: Adding constants

Suppose we replace every x by

$$x + 100$$

and every y by

$$y - 20.$$

The scatter plot shifts horizontally and vertically.
But the overall shape and linear trend do not change.

Skill 6: Adding constants

Suppose we replace every x by

$$x + 100$$

and every y by

$$y - 20.$$

The scatter plot shifts horizontally and vertically.
But the overall shape and linear trend do not change.
Therefore:

r does not change.

Why adding constants does not change r

Suppose:

$$x_i^{\text{new}} = x_i + k.$$

Why adding constants does not change r

Suppose:

$$x_i^{\text{new}} = x_i + k.$$

Then:

$$\bar{x}_{\text{new}} = \bar{x} + k.$$

Why adding constants does not change r

Suppose:

$$x_i^{\text{new}} = x_i + k.$$

Then:

$$\bar{x}_{\text{new}} = \bar{x} + k.$$

So:

$$x_i^{\text{new}} - \bar{x}_{\text{new}} = (x_i + k) - (\bar{x} + k) = x_i - \bar{x}.$$

Why adding constants does not change r

Suppose:

$$x_i^{\text{new}} = x_i + k.$$

Then:

$$\bar{x}_{\text{new}} = \bar{x} + k.$$

So:

$$x_i^{\text{new}} - \bar{x}_{\text{new}} = (x_i + k) - (\bar{x} + k) = x_i - \bar{x}.$$

All deviations from the mean stay the same.

Why adding constants does not change r

Suppose:

$$x_i^{\text{new}} = x_i + k.$$

Then:

$$\bar{x}_{\text{new}} = \bar{x} + k.$$

So:

$$x_i^{\text{new}} - \bar{x}_{\text{new}} = (x_i + k) - (\bar{x} + k) = x_i - \bar{x}.$$

All deviations from the mean stay the same.

Therefore covariance, standard deviation, and correlation are unchanged.

Example 6A: Shift both variables

Suppose the original data have

$$r = 0.62.$$

Now define:

$$X = x + 50, \quad Y = y - 7.$$

Example 6A: Shift both variables

Suppose the original data have

$$r = 0.62.$$

Now define:

$$X = x + 50, \quad Y = y - 7.$$

No calculation is needed.

$$r_{X,Y} = 0.62.$$

Example 6A: Shift both variables

Suppose the original data have

$$r = 0.62.$$

Now define:

$$X = x + 50, \quad Y = y - 7.$$

No calculation is needed.

$$\boxed{r_{X,Y} = 0.62}.$$

Adding constants changes location, not the pattern of association.

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

The axes are stretched.

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

The axes are stretched.

The shape of the scatter plot is unchanged.

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

The axes are stretched.

The shape of the scatter plot is unchanged.

Covariance changes, and each standard deviation changes.

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

The axes are stretched.

The shape of the scatter plot is unchanged.

Covariance changes, and each standard deviation changes.

But the changes cancel in

$$r = \frac{c_{xy}}{s_x s_y}.$$

Positive rescaling

Suppose:

$$X = 3x, \quad Y = 10y.$$

The axes are stretched.

The shape of the scatter plot is unchanged.

Covariance changes, and each standard deviation changes.

But the changes cancel in

$$r = \frac{c_{xy}}{s_x s_y}.$$

Therefore:

r is unchanged.

Example 6B: Convert units

Suppose height and weight have

$$r = 0.74.$$

Height is converted from inches to centimeters:

$$X = 2.54x.$$

Weight is converted from pounds to kilograms:

$$Y \approx 0.454y.$$

Example 6B: Convert units

Suppose height and weight have

$$r = 0.74.$$

Height is converted from inches to centimeters:

$$X = 2.54x.$$

Weight is converted from pounds to kilograms:

$$Y \approx 0.454y.$$

Both are positive rescalings.

Example 6B: Convert units

Suppose height and weight have

$$r = 0.74.$$

Height is converted from inches to centimeters:

$$X = 2.54x.$$

Weight is converted from pounds to kilograms:

$$Y \approx 0.454y.$$

Both are positive rescalings.

Therefore:

$$r = 0.74 \text{ still.}$$

Covariance versus correlation under rescaling

Suppose:

$$X = 3x, \quad Y = 2y.$$

Covariance versus correlation under rescaling

Suppose:

$$X = 3x, \quad Y = 2y.$$

Then covariance is multiplied by

$$3 \cdot 2 = 6.$$

Covariance versus correlation under rescaling

Suppose:

$$X = 3x, \quad Y = 2y.$$

Then covariance is multiplied by

$$3 \cdot 2 = 6.$$

Also:

$$s_X = 3s_x, \quad s_Y = 2s_y.$$

Covariance versus correlation under rescaling

Suppose:

$$X = 3x, \quad Y = 2y.$$

Then covariance is multiplied by

$$3 \cdot 2 = 6.$$

Also:

$$s_X = 3s_x, \quad s_Y = 2s_y.$$

So:

$$r_{X,Y} = \frac{6c_{xy}}{(3s_x)(2s_y)} = \frac{c_{xy}}{s_x s_y} = r_{x,y}.$$

Covariance versus correlation under rescaling

Suppose:

$$X = 3x, \quad Y = 2y.$$

Then covariance is multiplied by

$$3 \cdot 2 = 6.$$

Also:

$$s_X = 3s_x, \quad s_Y = 2s_y.$$

So:

$$r_{X,Y} = \frac{6c_{xy}}{(3s_x)(2s_y)} = \frac{c_{xy}}{s_x s_y} = r_{x,y}.$$

This is why correlation is standardized but covariance is not.

Skill 7: Correlation is not causation

A large value of $|r|$ tells us that two variables have a strong **linear association**.

Skill 7: Correlation is not causation

A large value of $|r|$ tells us that two variables have a strong **linear association**.
It does **not** prove:

x causes y .

Skill 7: Correlation is not causation

A large value of $|r|$ tells us that two variables have a strong **linear association**.
It does **not** prove:

x causes y .

Possible explanations include:

- ▶ x causes y ;
- ▶ y causes x ;
- ▶ another variable influences both;
- ▶ the association is partly or entirely accidental.

Example 7A: A hidden variable

Suppose ice-cream sales and swimming incidents both rise during the same months.

Example 7A: A hidden variable

Suppose ice-cream sales and swimming incidents both rise during the same months.
It would be unreasonable to conclude:

ice cream causes swimming incidents.

Example 7A: A hidden variable

Suppose ice-cream sales and swimming incidents both rise during the same months.
It would be unreasonable to conclude:

ice cream causes swimming incidents.

A plausible hidden variable is:

temperature or season.

Example 7A: A hidden variable

Suppose ice-cream sales and swimming incidents both rise during the same months. It would be unreasonable to conclude:

ice cream causes swimming incidents.

A plausible hidden variable is:

temperature or season.

Warmer weather can increase both variables.

Example 7B: Another hidden variable

Suppose larger fires tend to have:

- ▶ more firefighters present,
- ▶ more property damage.

Example 7B: Another hidden variable

Suppose larger fires tend to have:

- ▶ more firefighters present,
- ▶ more property damage.

A positive correlation between number of firefighters and damage does not imply that firefighters cause the damage.

Example 7B: Another hidden variable

Suppose larger fires tend to have:

- ▶ more firefighters present,
- ▶ more property damage.

A positive correlation between number of firefighters and damage does not imply that firefighters cause the damage.

A hidden variable is:

severity of the fire .

A useful causality checklist

When you see a strong correlation, ask:

- 1 Could x cause y ?

A useful causality checklist

When you see a strong correlation, ask:

- 1 Could x cause y ?
- 2 Could y cause x ?

A useful causality checklist

When you see a strong correlation, ask:

- 1 Could x cause y ?
- 2 Could y cause x ?
- 3 Could a third variable affect both?

A useful causality checklist

When you see a strong correlation, ask:

- 1 Could x cause y ?
- 2 Could y cause x ?
- 3 Could a third variable affect both?
- 4 Is there a scientific or logical mechanism?

A useful causality checklist

When you see a strong correlation, ask:

- ① Could x cause y ?
- ② Could y cause x ?
- ③ Could a third variable affect both?
- ④ Is there a scientific or logical mechanism?
- ⑤ Was the relationship observed experimentally or only observationally?

Skill 8: One point can strongly affect r

Correlation can be sensitive to unusual observations.

Skill 8: One point can strongly affect r

Correlation can be sensitive to unusual observations.

A point with an x -value far from the rest of the data can have especially large influence.

Skill 8: One point can strongly affect r

Correlation can be sensitive to unusual observations.

A point with an x -value far from the rest of the data can have especially large influence. Such a point is often called a **high-leverage point**.

Skill 8: One point can strongly affect r

Correlation can be sensitive to unusual observations.

A point with an x -value far from the rest of the data can have especially large influence. Such a point is often called a **high-leverage point**.

Therefore:

plot the data before and after investigating unusual points.

Example 8A: Without the distant point

Consider:

$(1, 2), (2, 5), (3, 3), (4, 6), (5, 4).$

Example 8A: Without the distant point

Consider:

$(1, 2), (2, 5), (3, 3), (4, 6), (5, 4).$

For these five points:

$$r = 0.50.$$

Example 8A: Without the distant point

Consider:

$(1, 2), (2, 5), (3, 3), (4, 6), (5, 4).$

For these five points:

$$r = 0.50.$$

That is only a moderate positive linear association.

Example 8A: Add one distant point

Now add:

$(20, 20)$.

Example 8A: Add one distant point

Now add:

$(20, 20)$.

The new correlation becomes approximately:

$$r = 0.979.$$

Example 8A: Add one distant point

Now add:

(20, 20).

The new correlation becomes approximately:

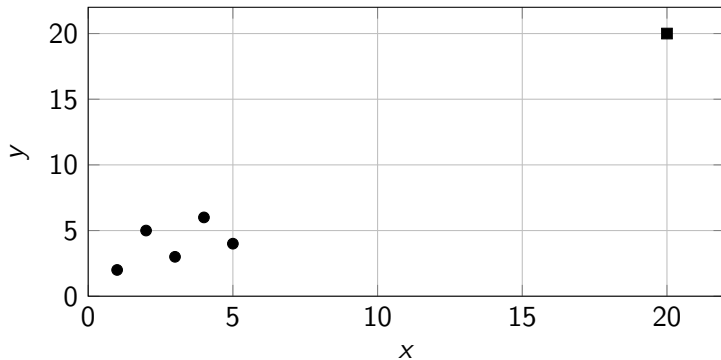
$$r = 0.979.$$

One distant point changes the numerical impression from moderate to extremely strong.

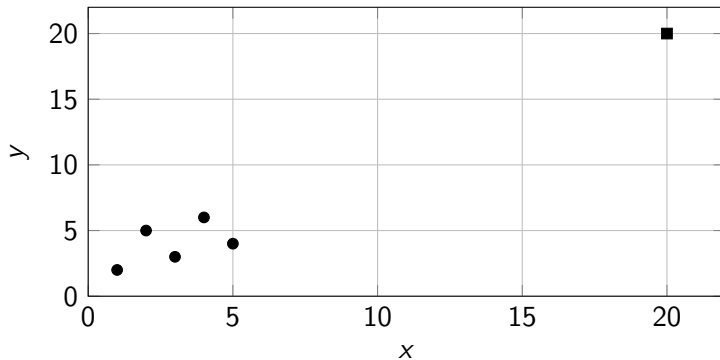
Example 8A: Visual comparison

The distant point exerts much more influence than an ordinary point inside the main cluster.

Example 8A: Visual comparison



Example 8A: Visual comparison



The distant point exerts much more influence than an ordinary point inside the main cluster.

Skill 9: Look for groups

A single correlation coefficient can hide important subgroup structure.

Skill 9: Look for groups

A single correlation coefficient can hide important subgroup structure.
Suppose the data come from two different populations.

Skill 9: Look for groups

A single correlation coefficient can hide important subgroup structure.
Suppose the data come from two different populations.
Each population may have a weak association internally.

Skill 9: Look for groups

A single correlation coefficient can hide important subgroup structure.
Suppose the data come from two different populations.
Each population may have a weak association internally.
But combining the groups can create an apparently strong overall association.

Example 9A: Two weak groups

Group A:

$(1, 2), (2, 1), (3, 3), (4, 2).$

Group B:

$(7, 8), (8, 7), (9, 9), (10, 8).$

Example 9A: Two weak groups

Group A:

$(1, 2), (2, 1), (3, 3), (4, 2).$

Group B:

$(7, 8), (8, 7), (9, 9), (10, 8).$

Within each group:

$$r \approx 0.316.$$

Example 9A: Two weak groups

Group A:

$(1, 2), (2, 1), (3, 3), (4, 2).$

Group B:

$(7, 8), (8, 7), (9, 9), (10, 8).$

Within each group:

$$r \approx 0.316.$$

So each subgroup has only a weak positive linear association.

Example 9A: Combine the groups

If we combine all eight points:

$$r \approx 0.937.$$

Example 9A: Combine the groups

If we combine all eight points:

$$r \approx 0.937.$$

That looks like a very strong positive correlation.

Example 9A: Combine the groups

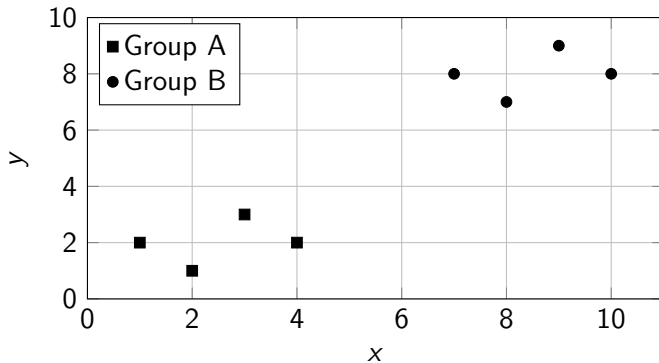
If we combine all eight points:

$$r \approx 0.937.$$

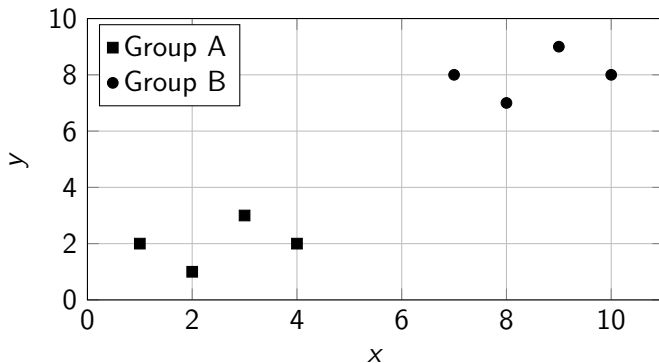
That looks like a very strong positive correlation.

But most of that apparent association comes from the **separation between the groups**, not from the relationship inside either group.

Example 9A: See the two clusters



Example 9A: See the two clusters



The combined r is not a good summary of the within-group relationships.

Skill 10: Predict r from context

Sometimes no calculation is needed.

Skill 10: Predict r from context

Sometimes no calculation is needed.

Example: adult height versus adult weight.

Would you expect:

- ① near $+1$,
- ② moderately positive,
- ③ close to 0 ,
- ④ moderately negative,
- ⑤ near -1 ?

Skill 10: Predict r from context

Sometimes no calculation is needed.

Example: adult height versus adult weight.

Would you expect:

- ① near $+1$,
- ② moderately positive,
- ③ close to 0,
- ④ moderately negative,
- ⑤ near -1 ?

A reasonable answer is:

moderately positive.

Taller adults tend to weigh more, but many other factors affect weight.

Example 10A: Car age and repair cost

A car is kept for ten years.

Each year we record:

x = age of the car,

y = repair and maintenance cost.

Example 10A: Car age and repair cost

A car is kept for ten years.

Each year we record:

x = age of the car,

y = repair and maintenance cost.

Older cars often require more repairs.

Example 10A: Car age and repair cost

A car is kept for ten years.

Each year we record:

x = age of the car,

y = repair and maintenance cost.

Older cars often require more repairs.

So we would generally expect:

$$r > 0.$$

Example 10A: Car age and repair cost

A car is kept for ten years.

Each year we record:

x = age of the car,

y = repair and maintenance cost.

Older cars often require more repairs.

So we would generally expect:

$$r > 0.$$

Probably positive but not perfect, because yearly repair costs can vary substantially.

Example 10B: Correct and incorrect answers

A ten-question exam records for each student:

x = number correct,

y = number incorrect.

Example 10B: Correct and incorrect answers

A ten-question exam records for each student:

x = number correct,

y = number incorrect.

For every student:

$$x + y = 10.$$

Example 10B: Correct and incorrect answers

A ten-question exam records for each student:

x = number correct,

y = number incorrect.

For every student:

$$x + y = 10.$$

So:

$$y = 10 - x.$$

Example 10B: Correct and incorrect answers

A ten-question exam records for each student:

x = number correct,

y = number incorrect.

For every student:

$$x + y = 10.$$

So:

$$y = 10 - x.$$

Every point lies exactly on a line with negative slope.

$$\boxed{r = -1}.$$

Mixed Example 1

A scatter plot shows points tightly clustered around a downward-sloping line.
Which value is most plausible?

0.92, 0.15, -0.18 , -0.91

Mixed Example 1

A scatter plot shows points tightly clustered around a downward-sloping line.
Which value is most plausible?

0.92, 0.15, -0.18 , -0.91

Downward means negative.

Mixed Example 1

A scatter plot shows points tightly clustered around a downward-sloping line.
Which value is most plausible?

0.92, 0.15, -0.18 , -0.91

Downward means negative.

Tightly clustered means close to 1 in magnitude.

Mixed Example 1

A scatter plot shows points tightly clustered around a downward-sloping line.
Which value is most plausible?

0.92, 0.15, -0.18, -0.91

Downward means negative.

Tightly clustered means close to 1 in magnitude.

Therefore:

-0.91.

Mixed Example 2

Suppose:

$$r = 0.68.$$

Every x -value is increased by 200.

Every y -value is multiplied by 5.

What is the new correlation?

Mixed Example 2

Suppose:

$$r = 0.68.$$

Every x -value is increased by 200.

Every y -value is multiplied by 5.

What is the new correlation?

Adding a constant does not change r .

Positive rescaling does not change r .

Mixed Example 2

Suppose:

$$r = 0.68.$$

Every x -value is increased by 200.

Every y -value is multiplied by 5.

What is the new correlation?

Adding a constant does not change r .

Positive rescaling does not change r .

Therefore:

$$r = 0.68.$$

Mixed Example 3

A data set has:

$$r = 0.03.$$

Can we conclude that x and y have no relationship?

Mixed Example 3

A data set has:

$$r = 0.03.$$

Can we conclude that x and y have no relationship?

No.

Mixed Example 3

A data set has:

$$r = 0.03.$$

Can we conclude that x and y have no relationship?

No.

We may conclude only that there is little or no **linear** association.

Mixed Example 3

A data set has:

$$r = 0.03.$$

Can we conclude that x and y have no relationship?

No.

We may conclude only that there is little or no **linear** association.
A curved relationship may still be present.

Mixed Example 4

A study finds:

$$r = 0.89$$

between two variables.

Can we conclude that increasing x will cause y to increase?

Mixed Example 4

A study finds:

$$r = 0.89$$

between two variables.

Can we conclude that increasing x will cause y to increase?

No.

Mixed Example 4

A study finds:

$$r = 0.89$$

between two variables.

Can we conclude that increasing x will cause y to increase?

No.

A high correlation gives evidence of a strong linear association.
It does not by itself establish cause and effect.

Mixed Example 5

Suppose:

$$c_{xy} = -24, \quad s_x = 4, \quad s_y = 8.$$

Find r .

Mixed Example 5

Suppose:

$$c_{xy} = -24, \quad s_x = 4, \quad s_y = 8.$$

Find r .

$$r = \frac{-24}{(4)(8)}$$

Mixed Example 5

Suppose:

$$c_{xy} = -24, \quad s_x = 4, \quad s_y = 8.$$

Find r .

$$r = \frac{-24}{(4)(8)}$$

$$r = \frac{-24}{32}$$

Mixed Example 5

Suppose:

$$c_{xy} = -24, \quad s_x = 4, \quad s_y = 8.$$

Find r .

$$r = \frac{-24}{(4)(8)}$$

$$r = \frac{-24}{32}$$

$$r = -0.75.$$

Mixed Example 5

Suppose:

$$c_{xy} = -24, \quad s_x = 4, \quad s_y = 8.$$

Find r .

$$r = \frac{-24}{(4)(8)}$$

$$r = \frac{-24}{32}$$

$$\boxed{r = -0.75}.$$

This indicates a strong negative linear association.

Mixed Example 6

Suppose a scatter plot is strongly U-shaped.
Which is more plausible?

$$r = 0.95, \quad r = 0.02, \quad r = -0.90.$$

Mixed Example 6

Suppose a scatter plot is strongly U-shaped.
Which is more plausible?

$$r = 0.95, \quad r = 0.02, \quad r = -0.90.$$

A U-shape is not linear.

Mixed Example 6

Suppose a scatter plot is strongly U-shaped.
Which is more plausible?

$$r = 0.95, \quad r = 0.02, \quad r = -0.90.$$

A U-shape is not linear.

The upward tendency on one side can cancel the downward tendency on the other.

Mixed Example 6

Suppose a scatter plot is strongly U-shaped.
Which is more plausible?

$$r = 0.95, \quad r = 0.02, \quad r = -0.90.$$

A U-shape is not linear.

The upward tendency on one side can cancel the downward tendency on the other.
So:

$$r = 0.02$$

is entirely plausible.

The chapter in one slide

$$c_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

The chapter in one slide

$$c_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{c_{xy}}{s_x s_y}$$

The chapter in one slide

$$c_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{c_{xy}}{s_x s_y}$$

$$-1 \leq r \leq 1$$

The chapter in one slide

$$c_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{c_{xy}}{s_x s_y}$$

$$-1 \leq r \leq 1$$

$r \approx 1$: strong positive linear association

$r \approx -1$: strong negative linear association

$r \approx 0$: weak or no linear association

The chapter in one slide

$$c_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$r = \frac{c_{xy}}{s_x s_y}$$

$$-1 \leq r \leq 1$$

$r \approx 1$: strong positive linear association

$r \approx -1$: strong negative linear association

$r \approx 0$: weak or no linear association

And:

correlation does not imply causation.

A practical decision tree

When facing a bivariate-data problem:

- 1 Identify the paired variables x and y .

A practical decision tree

When facing a bivariate-data problem:

- ① Identify the paired variables x and y .
- ② Plot the data or inspect the scatter plot.

A practical decision tree

When facing a bivariate-data problem:

- 1 Identify the paired variables x and y .
- 2 Plot the data or inspect the scatter plot.
- 3 Describe direction, strength, and form.

A practical decision tree

When facing a bivariate-data problem:

- ① Identify the paired variables x and y .
- ② Plot the data or inspect the scatter plot.
- ③ Describe direction, strength, and form.
- ④ Compute r if a numerical measure is needed.

A practical decision tree

When facing a bivariate-data problem:

- ① Identify the paired variables x and y .
- ② Plot the data or inspect the scatter plot.
- ③ Describe direction, strength, and form.
- ④ Compute r if a numerical measure is needed.
- ⑤ Check whether the relationship is actually linear.

A practical decision tree

When facing a bivariate-data problem:

- ① Identify the paired variables x and y .
- ② Plot the data or inspect the scatter plot.
- ③ Describe direction, strength, and form.
- ④ Compute r if a numerical measure is needed.
- ⑤ Check whether the relationship is actually linear.
- ⑥ Look for unusual points or subgroups.

A practical decision tree

When facing a bivariate-data problem:

- ① Identify the paired variables x and y .
- ② Plot the data or inspect the scatter plot.
- ③ Describe direction, strength, and form.
- ④ Compute r if a numerical measure is needed.
- ⑤ Check whether the relationship is actually linear.
- ⑥ Look for unusual points or subgroups.
- ⑦ Do not claim causation from correlation alone.

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?
- ▶ Did I distinguish positive from negative association?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?
- ▶ Did I distinguish positive from negative association?
- ▶ Did I remember that $-1 \leq r \leq 1$?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?
- ▶ Did I distinguish positive from negative association?
- ▶ Did I remember that $-1 \leq r \leq 1$?
- ▶ Did I interpret r as a measure of **linear** association?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?
- ▶ Did I distinguish positive from negative association?
- ▶ Did I remember that $-1 \leq r \leq 1$?
- ▶ Did I interpret r as a measure of **linear** association?
- ▶ Did I inspect the scatter plot for curves, clusters, or unusual points?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I keep the x - and y -values paired correctly?
- ▶ Did I put x on the horizontal axis and y on the vertical axis?
- ▶ Did I distinguish positive from negative association?
- ▶ Did I remember that $-1 \leq r \leq 1$?
- ▶ Did I interpret r as a measure of **linear** association?
- ▶ Did I inspect the scatter plot for curves, clusters, or unusual points?
- ▶ Did I avoid claiming causation without additional evidence?

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Find the correlation coefficient:

$$r = \frac{15}{(2.5)(8)}.$$

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Find the correlation coefficient:

$$r = \frac{15}{(2.5)(8)}.$$

$$r = \frac{15}{20} = \boxed{0.75}.$$

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Find the correlation coefficient:

$$r = \frac{15}{(2.5)(8)}.$$

$$r = \frac{15}{20} = \boxed{0.75}.$$

Interpret:

strong positive linear association.

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Find the correlation coefficient:

$$r = \frac{15}{(2.5)(8)}.$$

$$r = \frac{15}{20} = \boxed{0.75}.$$

Interpret:

strong positive linear association.

Can we conclude that x causes y ?

Final practice

Suppose:

$$c_{xy} = 15, \quad s_x = 2.5, \quad s_y = 8.$$

Find the correlation coefficient:

$$r = \frac{15}{(2.5)(8)}.$$

$$r = \frac{15}{20} = \boxed{0.75}.$$

Interpret:

strong positive linear association.

Can we conclude that x causes y ?

No.