Relationships between variables.

- ***Association***

  Examples:

  - Smoking is associated with heart disease.
  - Weight is associated with height.
  - Income is associated with education.

- ***Functional relationships*** between quantitative variables. These allow us to predict the (unobserved) value of one variable based on the (observed) value of another. This goes beyond association and implies ***causation***. I.e., changes in the values of one variable *cause* the value of the other variable to change.

- Statistical studies can only ever determine *association* between variables. Determining a causal relationship requires a different type of study.

Example: The data in the table below is the *shoe-size/height* data from a sample of 18 high school students.

s	h		s	h
5	63		7	61
4	60		6.5	64
12	77		9	72
8	66		4	65
9	70		8	69
7.5	65		4	62
6.5	65		6	66
11.5	67		10.5	71
10.5	74		11	71

Summary Statistics:

$$\overline{s} = \frac{140}{18} \approx 7.77, \; SD_s \approx 2.58;$$

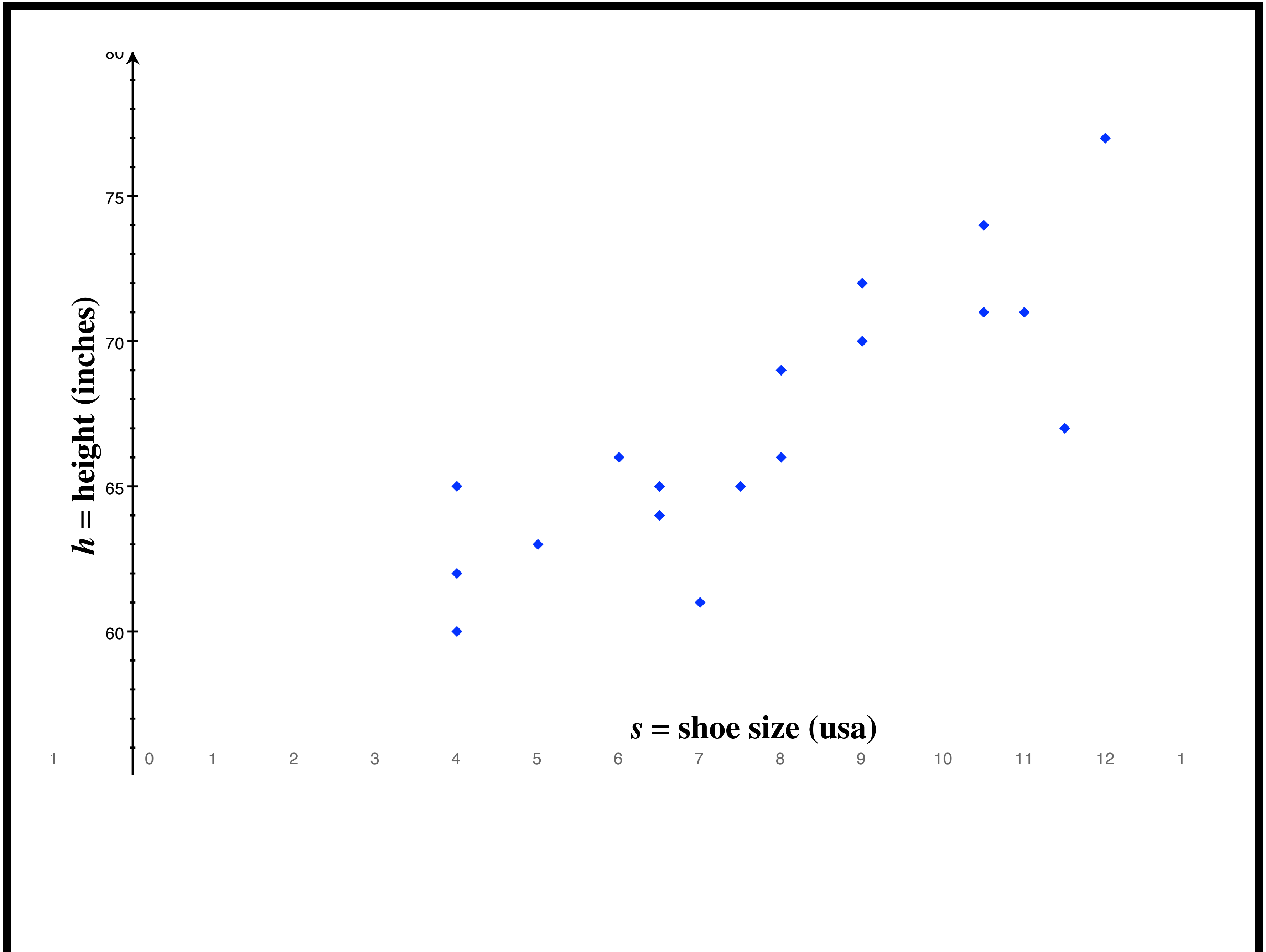
$$\overline{h} = \frac{1208}{18} \approx 67.11, \; SD_h \approx 4.54.$$

We can also represent this data as a set of pairs of values, as below:

$$\{(5,63),\quad (7,61),\quad (4,60),\quad (6.5,64),\quad (12,77),\quad (9,72),$$
$$(8,66),\quad (4,65),\quad (9,70),\quad (8,69),\quad (7.5,65),\quad (4,62),$$
$$(6.5,65),\quad (6,66),\quad (11.5,67),\quad (10.5,71),\quad (10.5,74),\quad (11,71)\}$$

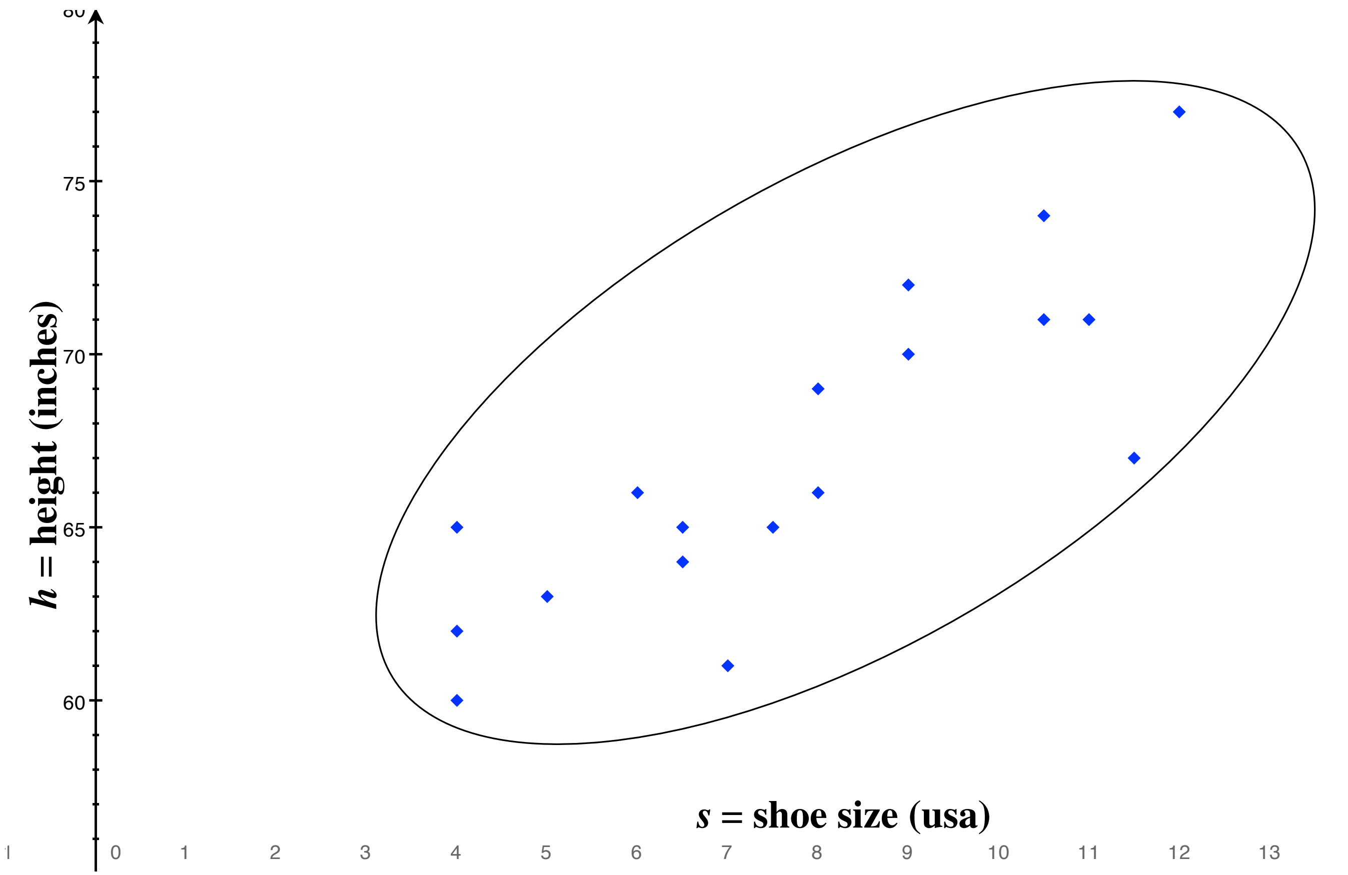
Important: The two coordinates of each pair *come from the same observation.*

(*) Paired data may be plotted as points in a 2-dimensional coordinate system. This type of plot is called a ***scatter plot***.



80
75
70
65
60
h = height (inches)
s = shoe size (usa)
0
1
2
3
4
5
6
7
8
9
10
11
12
1

The same scatter plot framed by an oval:

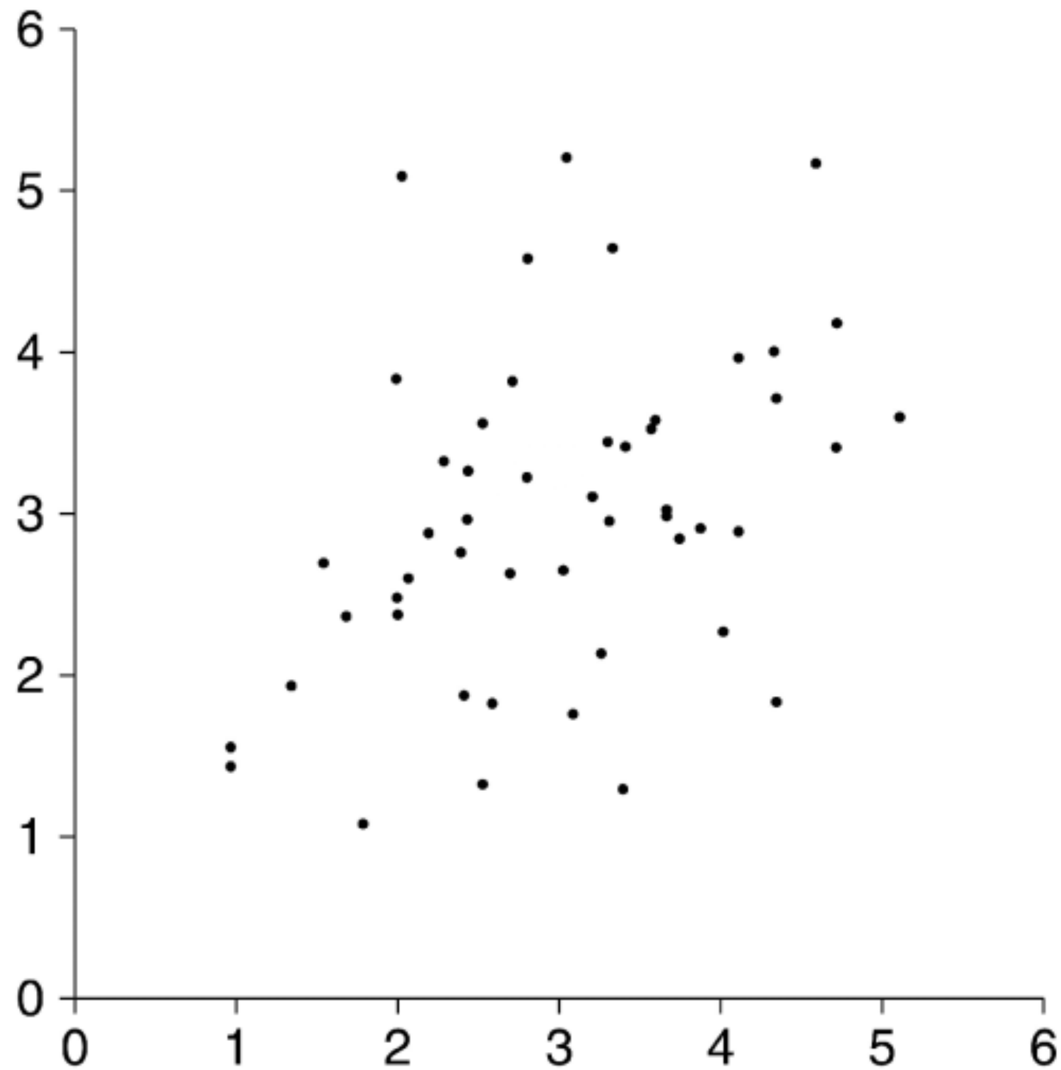


The direction of the oval indicates a ***positive*** relationship between shoe size and height. On average, people with bigger feet are taller than people with smaller feet.

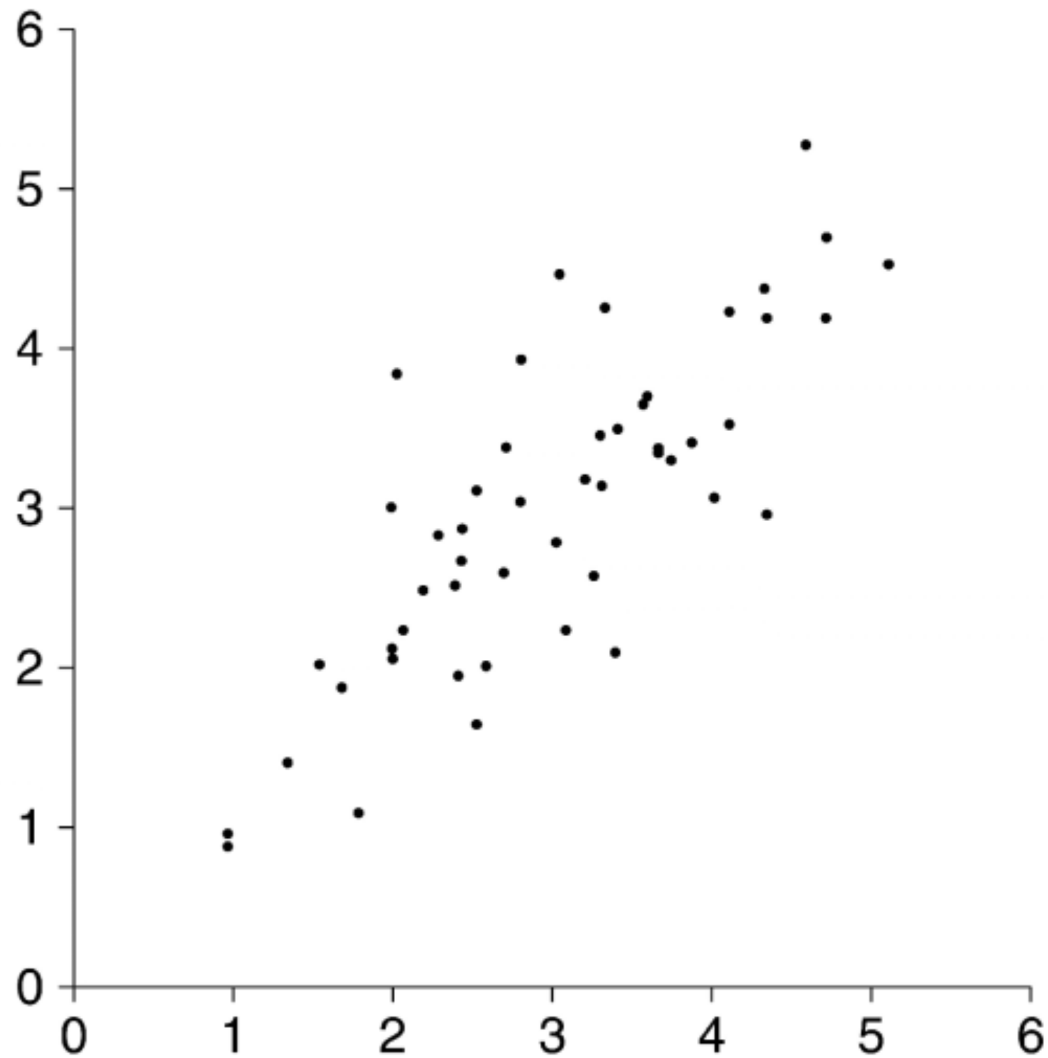
In general: the ‘shape’ of the scatter plot may give an indication of the type of relationship that might exist between the variables.

- *Positive:* y tends to get bigger when x is bigger.
- *Negative:* y tends to get smaller when x is bigger.
- *linear:* the points (x, y) in the scatterplot seem to cluster around a straight line.

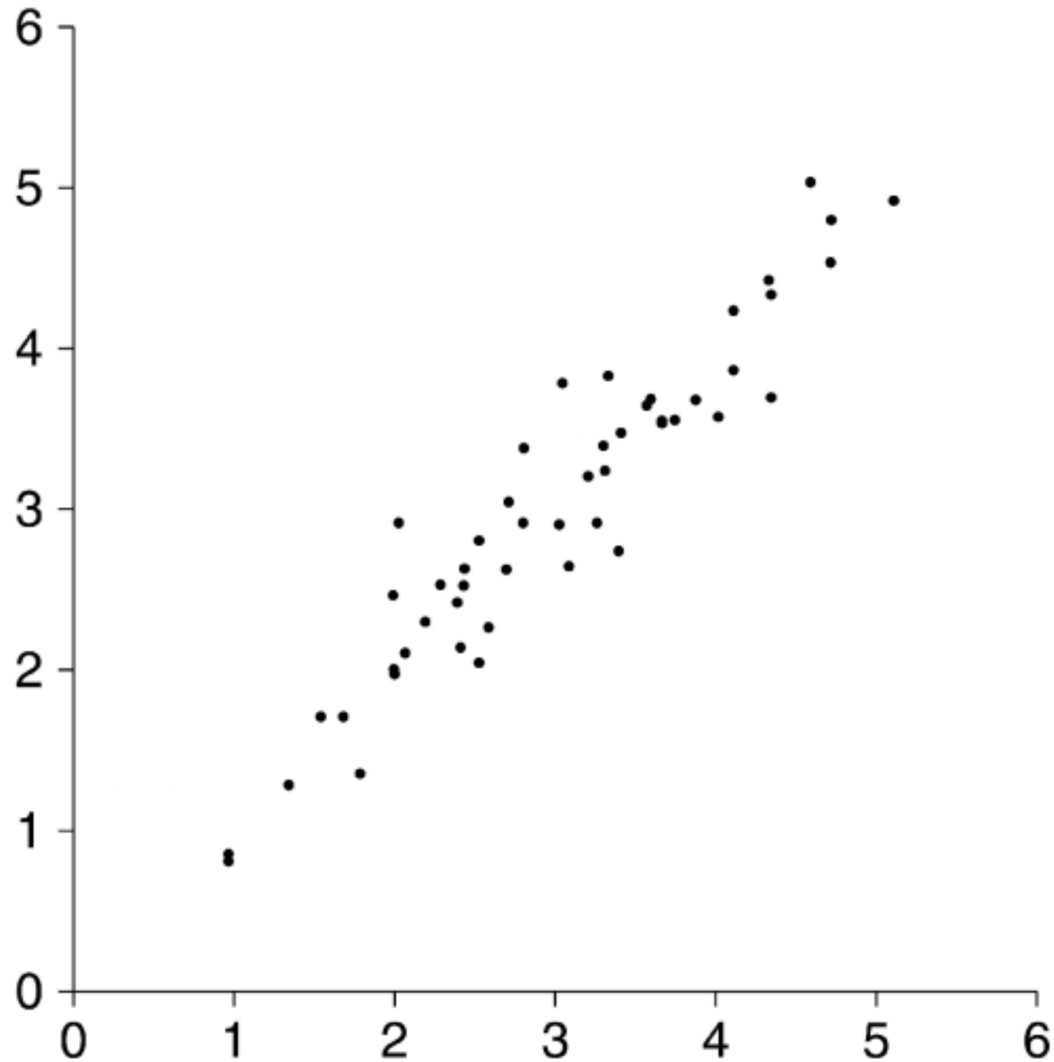
Observation: More complicated relationships can and do exist between variables. We are presently only considering the simplest ones.



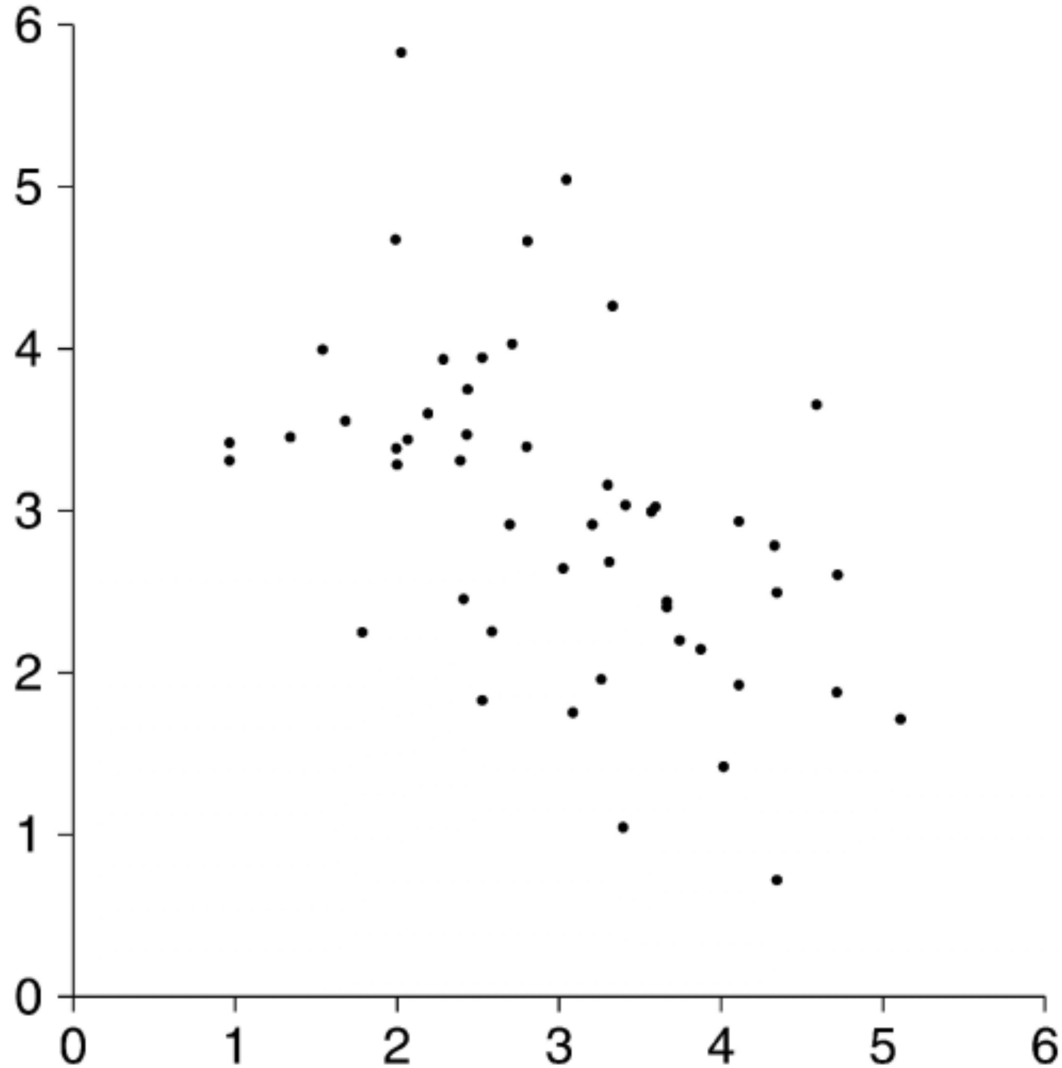
Weak positive association



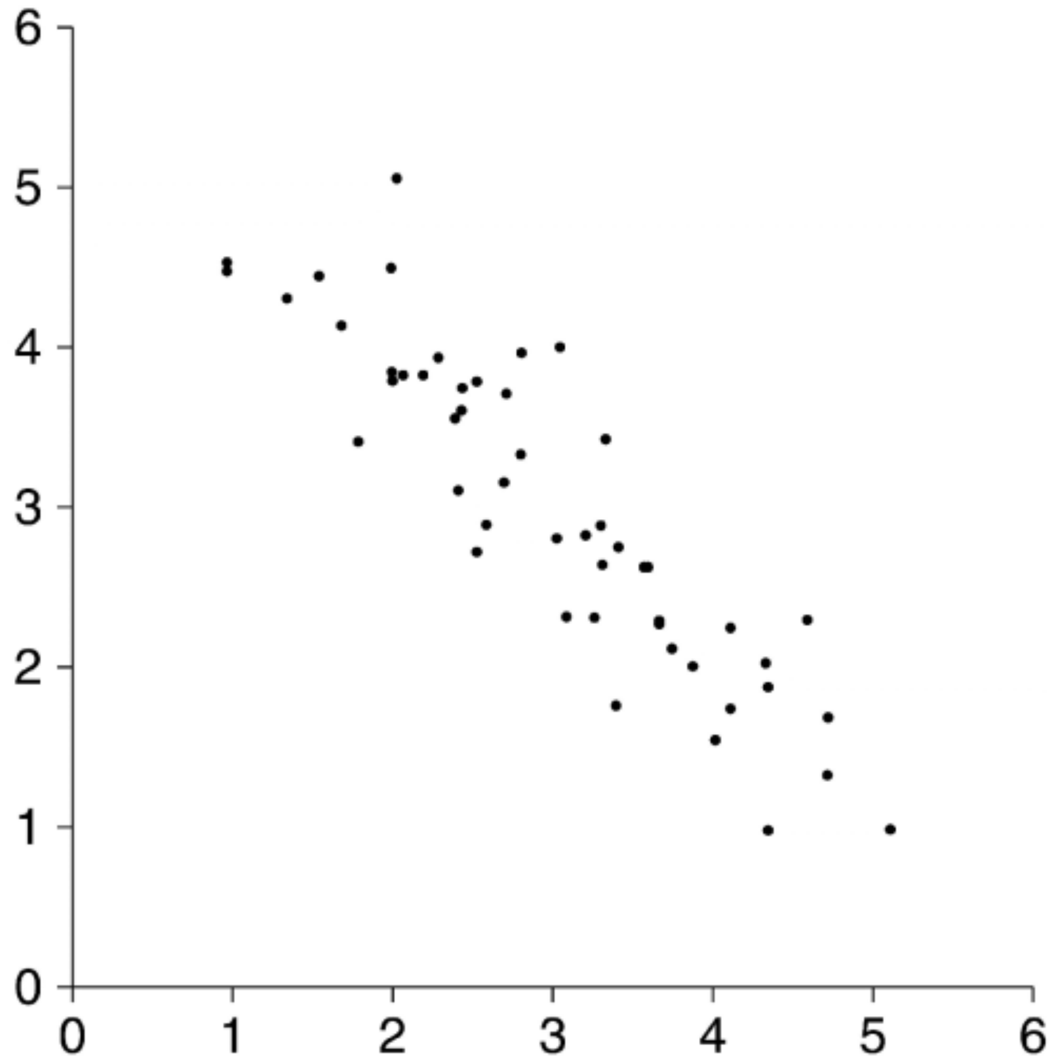
Stronger positive association



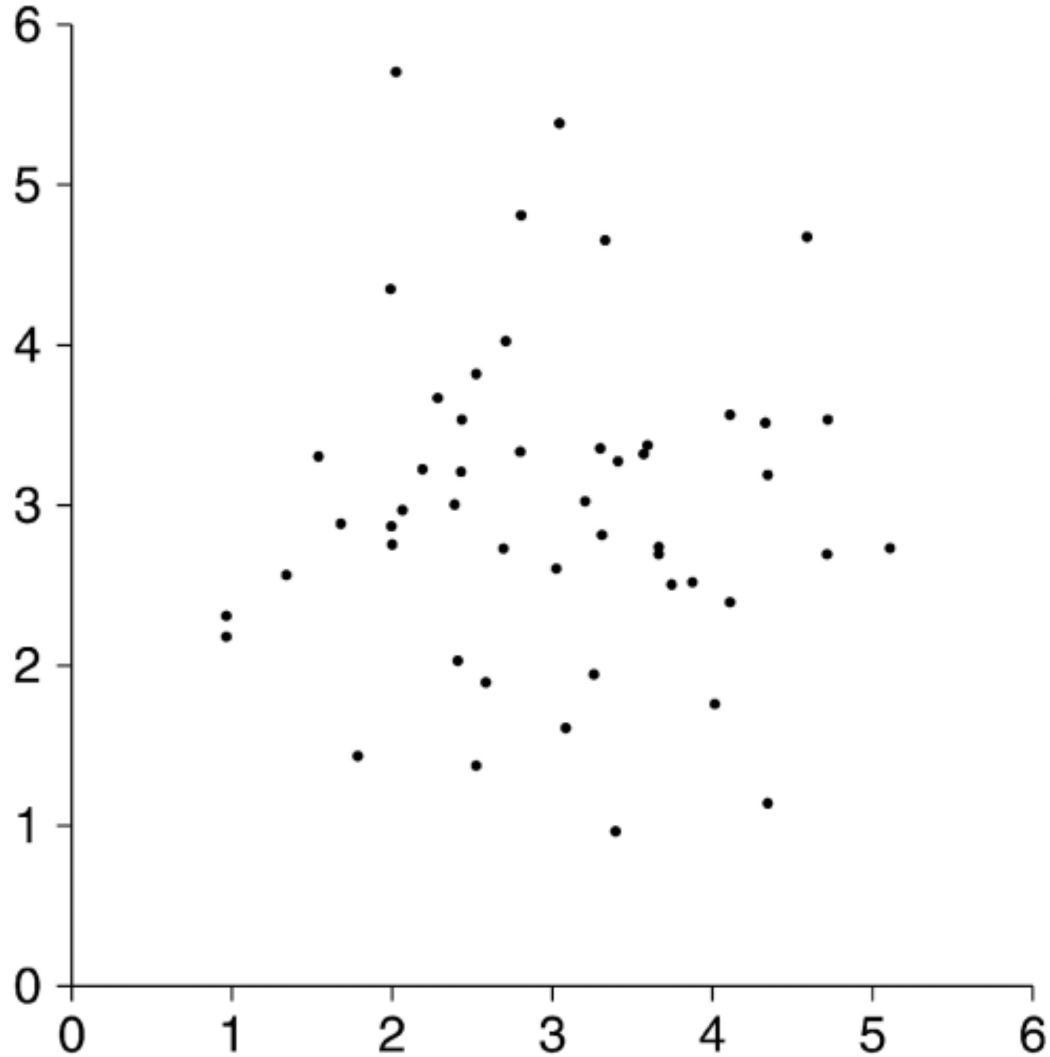
Very strong positive association



Weak negative association

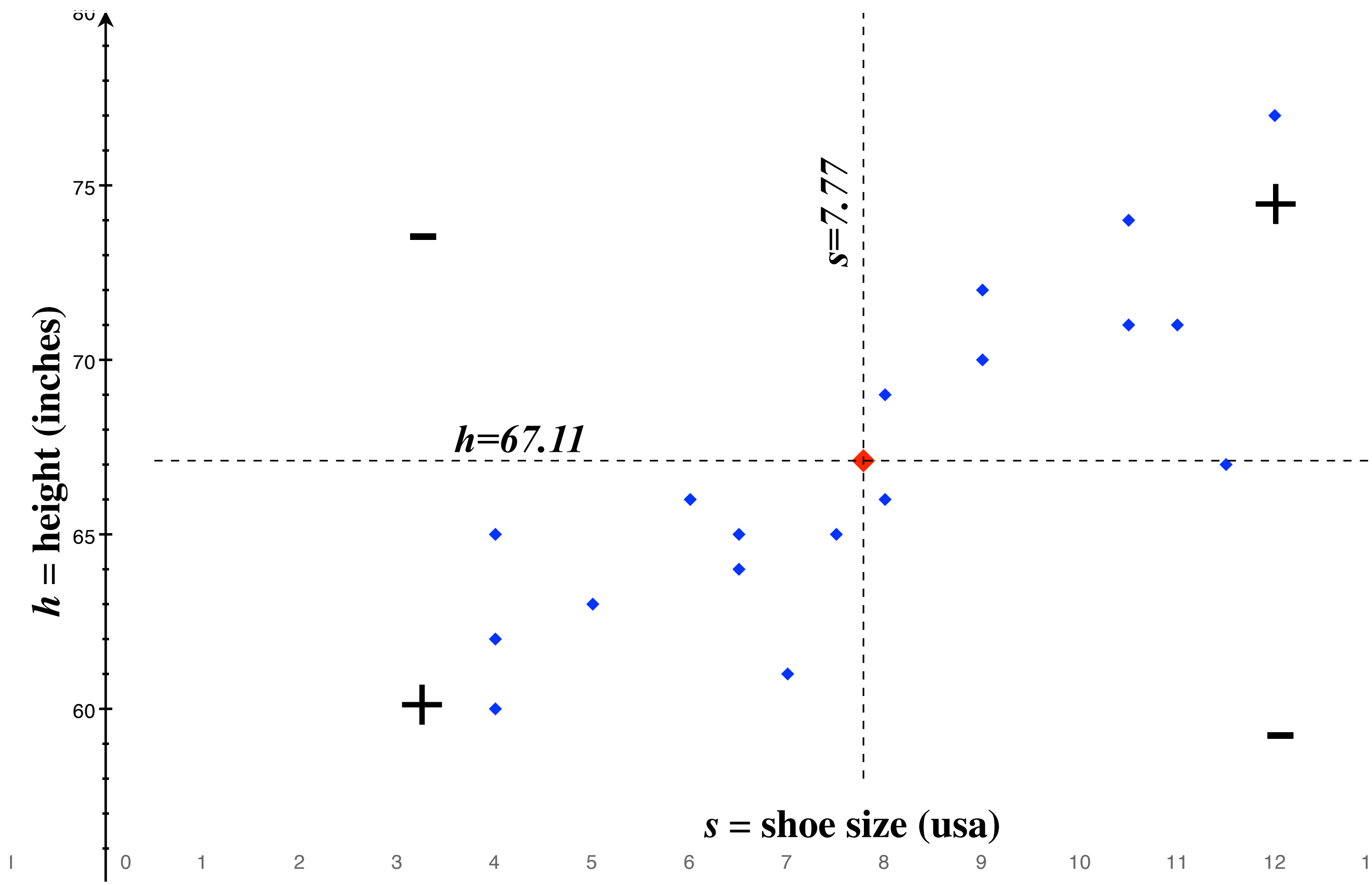


Strong negative association



No obvious (linear) relationship

The shoe size – height scatterplot with the point of averages (red diamond) and positive and negative quadrants.



Covariance

The ***covariance*** of the paired data

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$$

is given by the formula

$$\operatorname{cov}(x, y) = \frac{1}{n} \sum_{j=1}^{n} \left[(x_j - \overline{x}) \cdot (y_j - \overline{y}) \right],$$

where $\overline{x}$ is the average of $\{x_1, x_2, \ldots, x_n\}$ and $\overline{y}$ is the average of $\{y_1, y_2, \ldots, y_n\}$.

Observation: Points in the scatterplot that lie in the ***positive*** quadrants (see the figure on the previous page) contribute positive terms to the covariance sum, and points that lie in the ***negative*** quadrants contribute negative terms to the covariance sum.

Therefore...

- If $\text{cov}(x, y) > 0$, then the relationship between x and y is generally ***positive***.
- If $\text{cov}(x, y) < 0$, then the relationship between x and y is generally ***negative***.
- If $\text{cov}(x, y) = 0$, then we draw no conclusion.

Comment: The covariance is a good tool for detecting ***linear*** relationships. Two variables may have a very distinct *nonlinear* relationship, with zero covariance.

The scatterplot for shoe size vs. height suggests that $\text{cov}(s, h)$ will be positive. We can check this with a simple calculation:

$$\begin{aligned}\text{cov}(s,h) &= \frac{1}{18}\sum_{j=1}^{18}\left(s_j - \frac{70}{9}\right)\left(h_j - \frac{604}{9}\right)\\ &= \frac{1}{18}\left[\left(5 - \frac{70}{9}\right)\left(63 - \frac{604}{9}\right) + \cdots + \left(11 - \frac{70}{9}\right)\left(71 - \frac{604}{9}\right)\right]\\ &\approx 9.58 > 0.\end{aligned}$$

- $cov(s, h) > 0$ as expected, indicating a positive relationship.
- What does the *size* (9.58) of the covariance tell us about the relationship?
- Does larger covariance indicate a stronger relationship, or something else?

No... Covariance is sensitive to *changes of scale.*

If we measure the heights in cm instead of inches, we get a new variable $c_j = 2.54 \cdot h_j$ (because there are 2.54 cm to an inch). Moreover,

$$\overline{c} = \frac{1}{18}\sum_{j=1}^{18}(2.54 \cdot h_j) = 2.54 \cdot \left(\frac{1}{18}\sum_{j=1}^{18} h_j\right) = 2.54 \cdot \overline{h}.$$

If you calculate the covariance $\mathrm{cov}(s, c)$, you will find that

$$\mathrm{cov}(s, c) = 2.54 \cdot \mathrm{cov}(s, h)$$

(*) The biometric relationship between height and shoe size doesn't change depending on the units of height, but the covariance does.

(*) The ***sign*** of the covariance gave us useful information about the relationship but the ***size*** of the covariance, by itself, does not.

The correlation coefficient.

Given paired data, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, the ***correlation coefficient*** r_{xy} is defined by

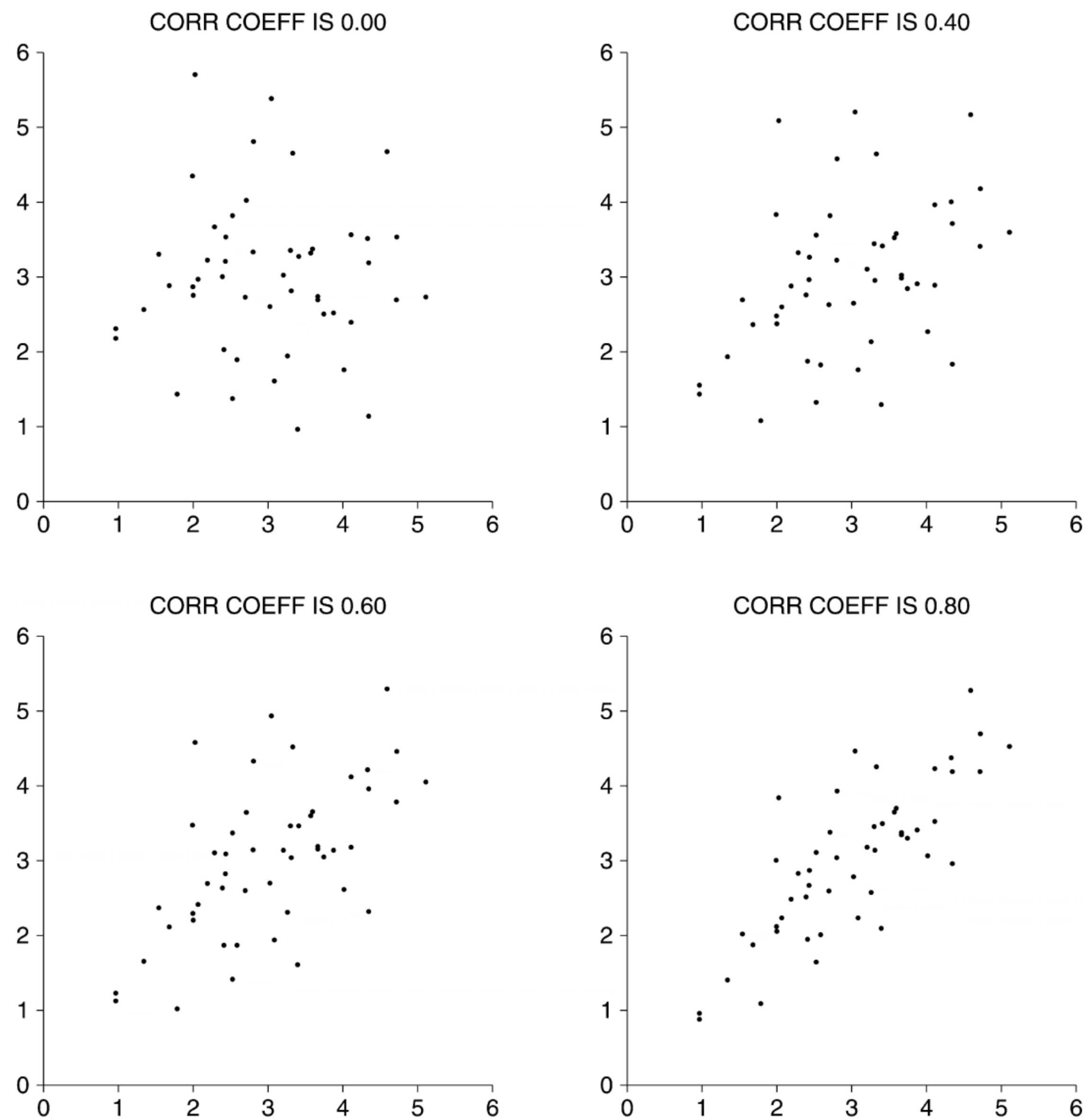
$$r_{xy} = \frac{\mathrm{cov}(x,y)}{SD_x \cdot SD_y} = \frac{1}{n}\sum_j \left(\frac{x_j - \overline{x}}{SD_x}\right) \cdot \left(\frac{y_j - \overline{y}}{SD_y}\right).$$

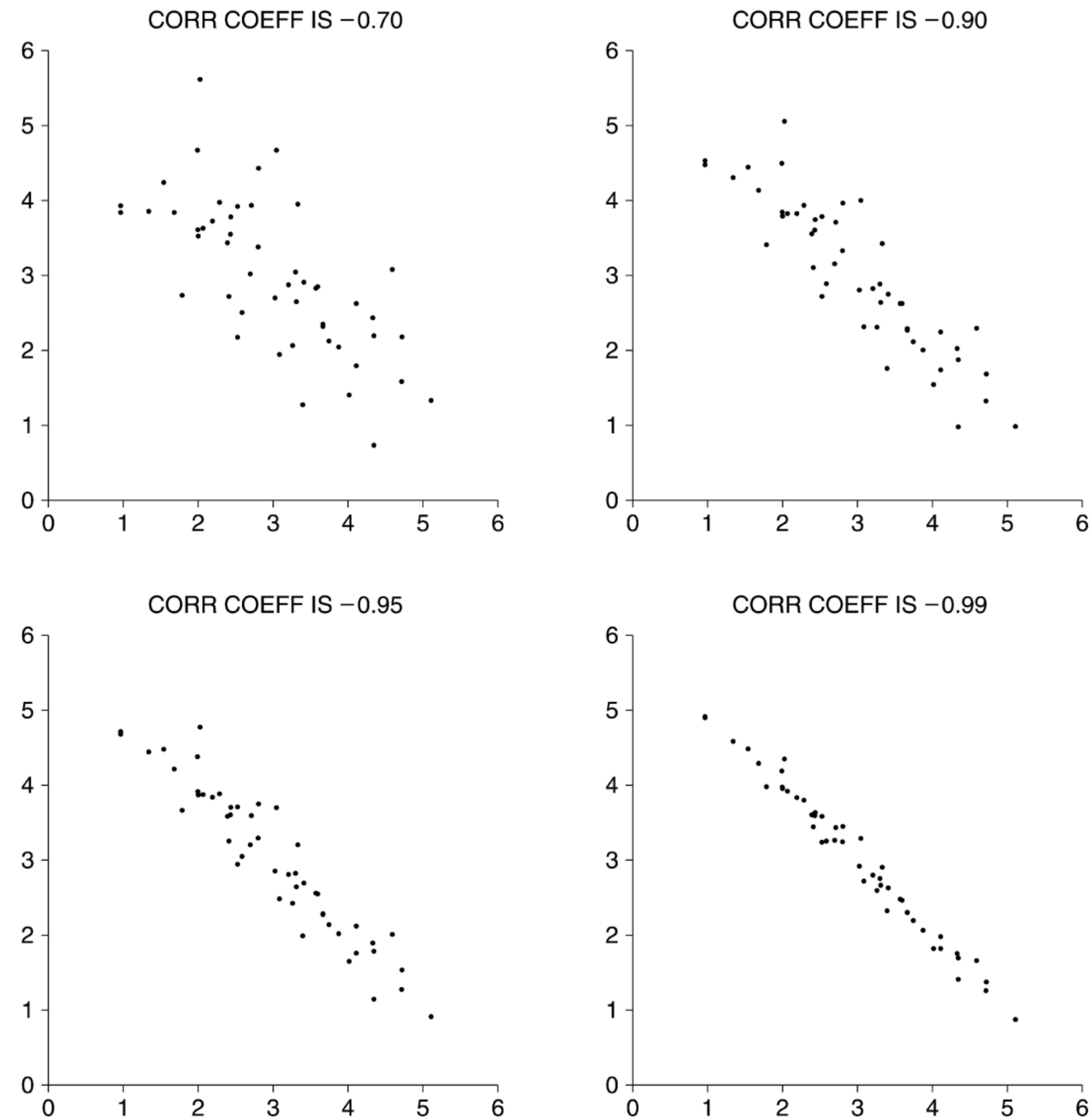
Observation: $\frac{x_j - \overline{x}}{SD_x} = z_{x_j}$ is the z-score of x_j and $\frac{y_j - \overline{y}}{SD_y} = z_{y_j}$ is the z-score of y_j. So

$$r_{xy} = \frac{1}{n}\sum_j z_{x_j} \cdot z_{y_j}.$$

Returning to the height/shoe size example, we have:

$$r_{sh} \approx \frac{9.58}{2.58 \cdot 4.54} \approx 0.818.$$


CORR COEFF IS 0.00
CORR COEFF IS 0.40
CORR COEFF IS 0.60
CORR COEFF IS 0.80
0
1
2
3
4
5
6


CORR COEFF IS −0.70
CORR COEFF IS −0.90
CORR COEFF IS −0.95
CORR COEFF IS −0.99
0
1
2
3
4
5
6

Properties of the correlation coefficient.

- r_{xy} is always between -1 and 1 (and is not sensitive to scale).
- If $r_{xy} > 0$, then there is a positive association between x and y.
- If $r_{xy} < 0$, then there is a negative association between x and y.
- The closer $|r_{xy}|$ is to 1, the stronger the (linear) association between the two variables. The closer r_{xy} is to 0, the weaker the (linear) association between the two variables.

Question: ***If there is strong correlation between the variables x and y (big $|r|$), what does this tell us about any causal relation between the variables?***

Answer: ***None*** by itself.

The correlation coefficient is a measure of statistical (linear) ***association***. It does not indicate ***causation***. In many cases where there is strong correlation, there are also significant confounding variables.

Examples.

- Shoe size and reading ability.
- Education level and unemployment.
- Range and duration of species.