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:

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

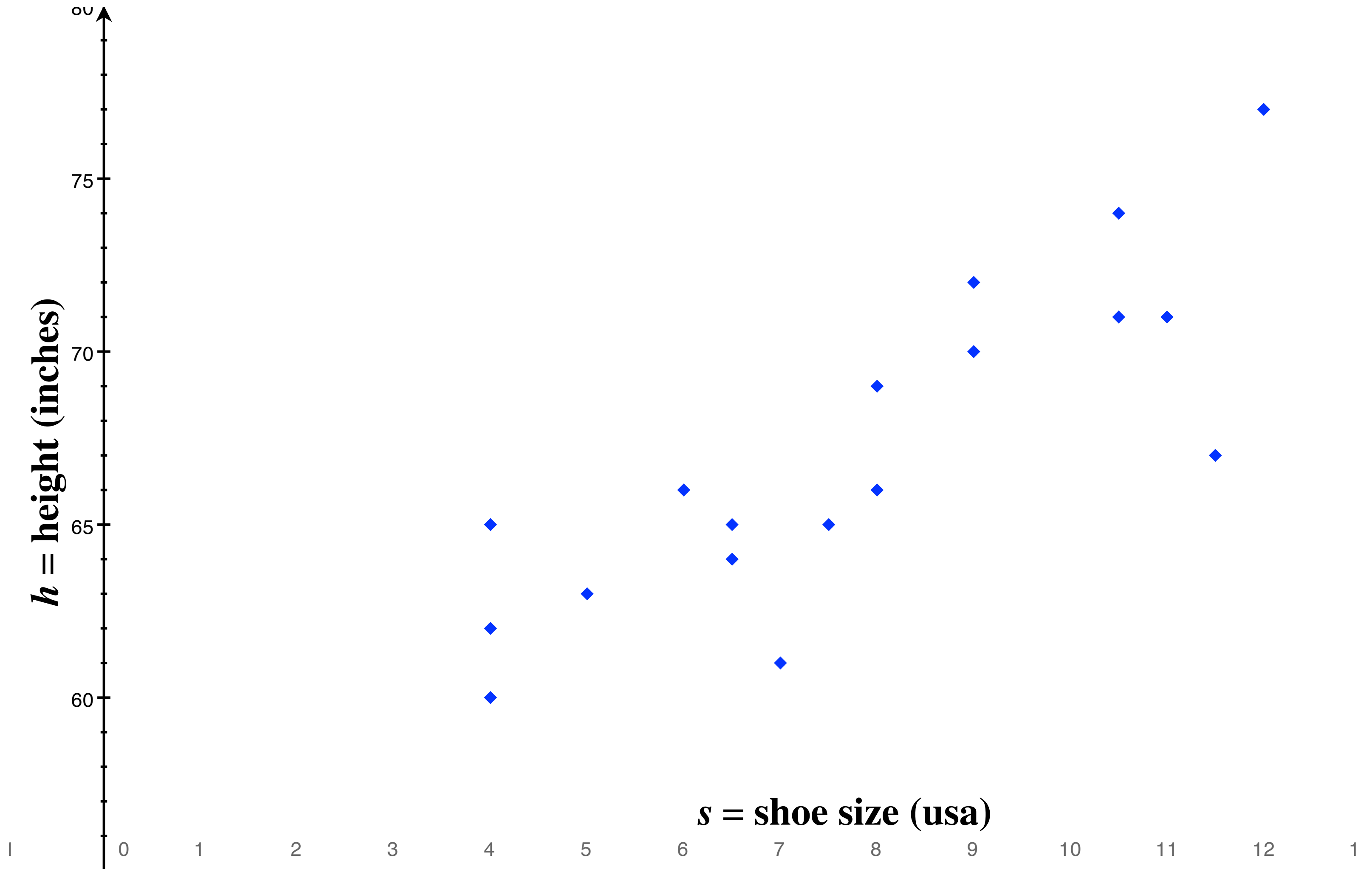
$$\overline{h} = \frac{1208}{18} \approx 67.11,\ SD_h \approx 4.54.$$

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

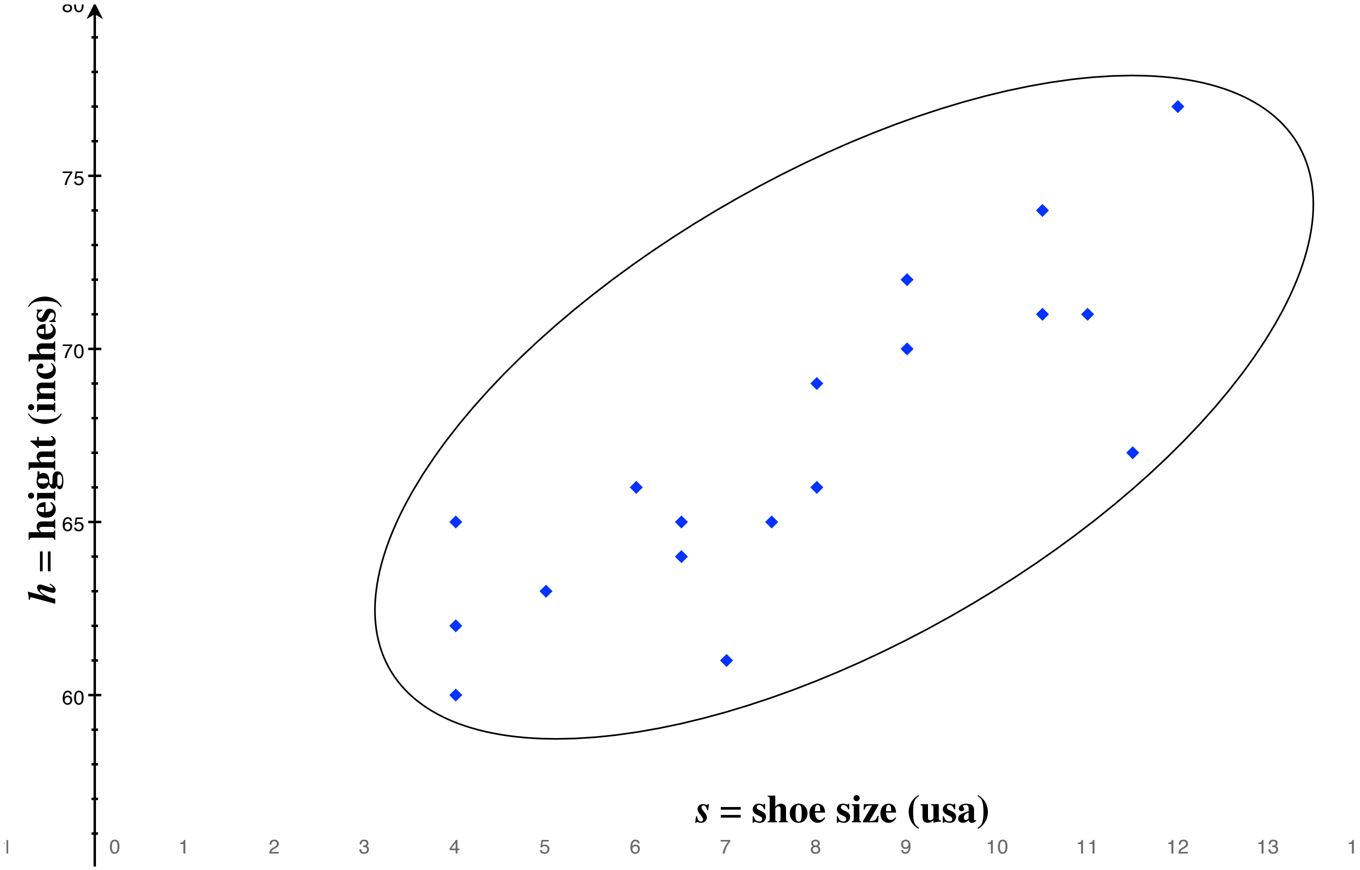
$$\begin{array}{cccccc}
\{\,(5,63), & (7,61), & (4,60), & (6.5,64), & (12,77), & (9,72), \\
(8,66), & (4,65), & (9,70), & (8,69), & (7.5,65), & (4,62), \\
(6.5,65), & (6,66), & (11.5,67), & (10.5,71), & (10.5,74), & (11,71)\,\}
\end{array}$$

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 is called a ***scatter plot***.


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

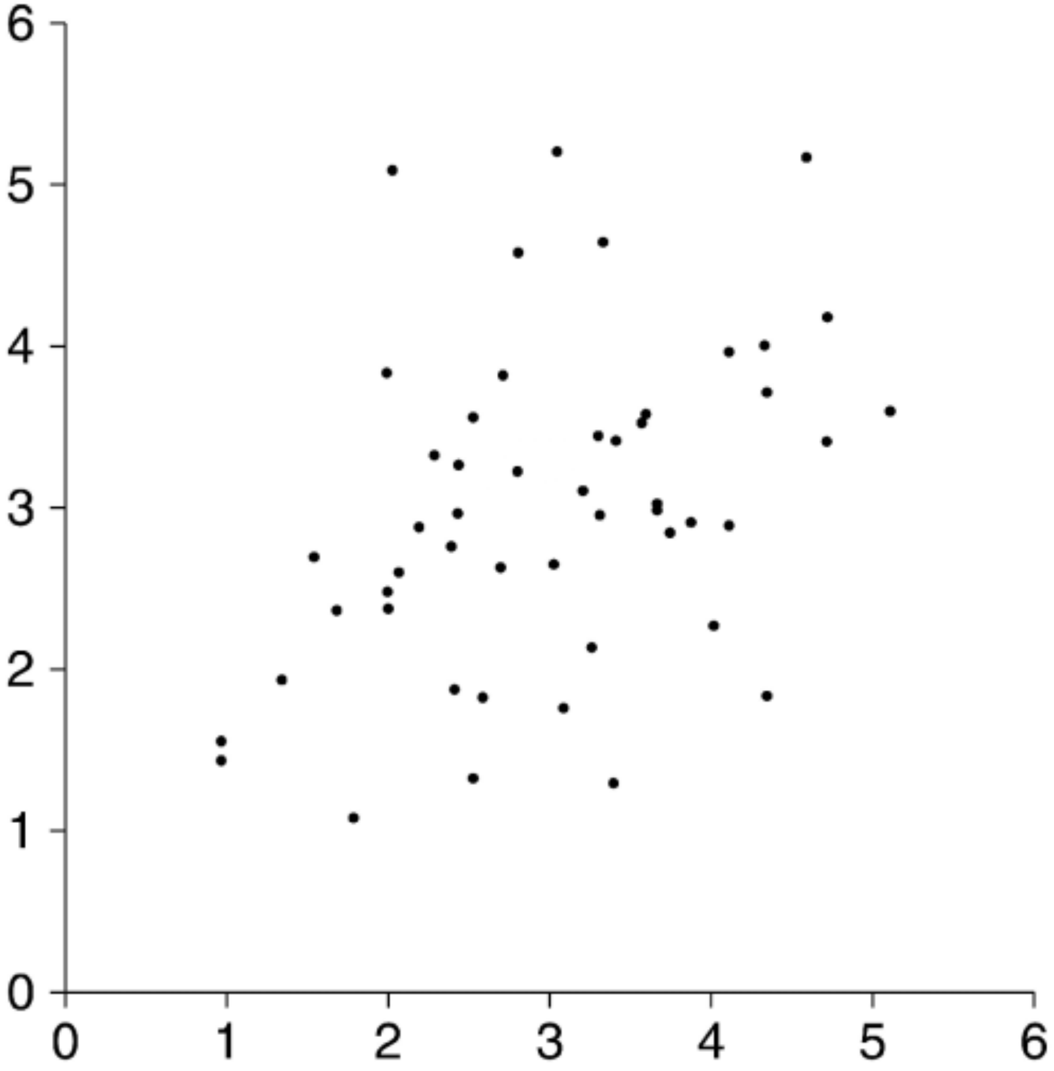
The same scatter plot framed by an oval:



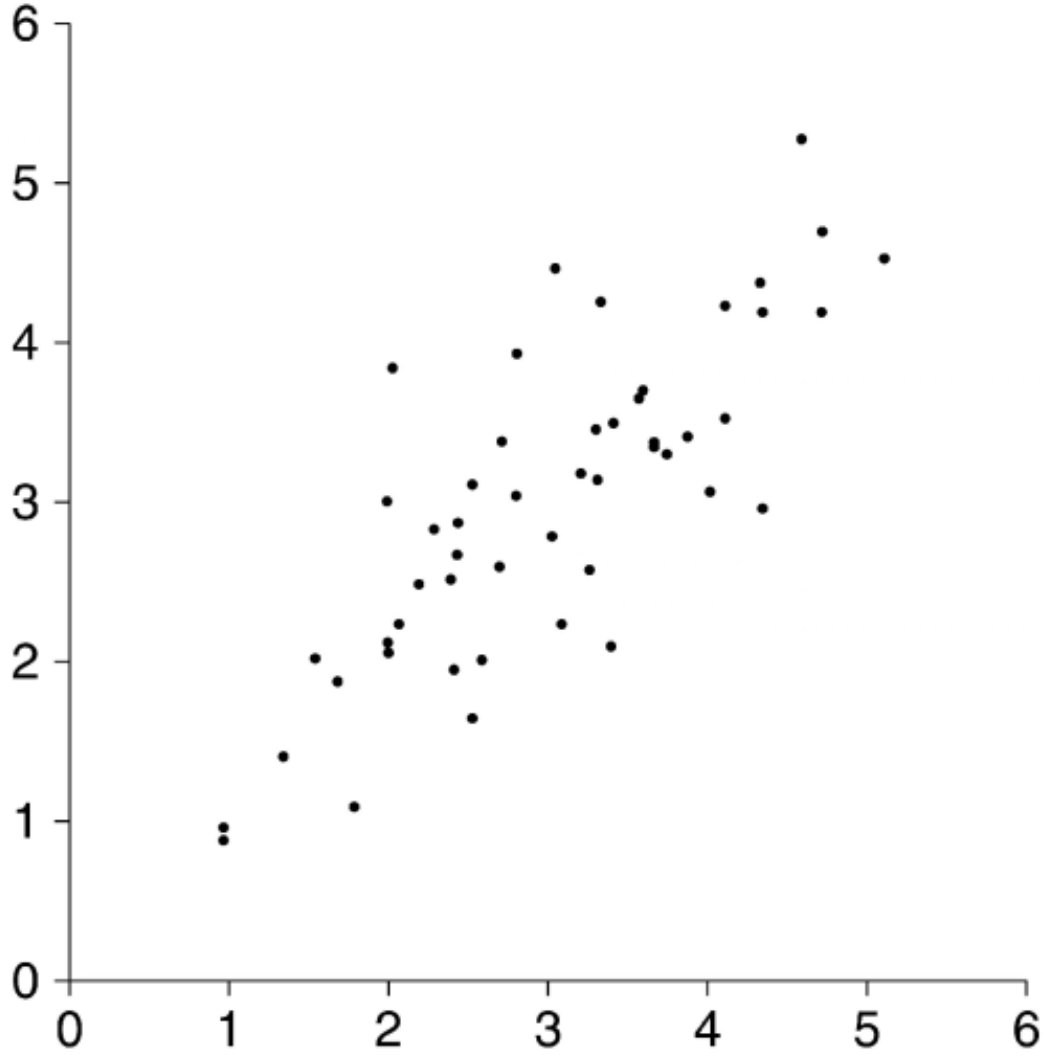
The direction of the oval indicates a ***positive*** relationship between shoe size and height. ***On average***, people with bigger feet are taller than people with smaller feet.

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

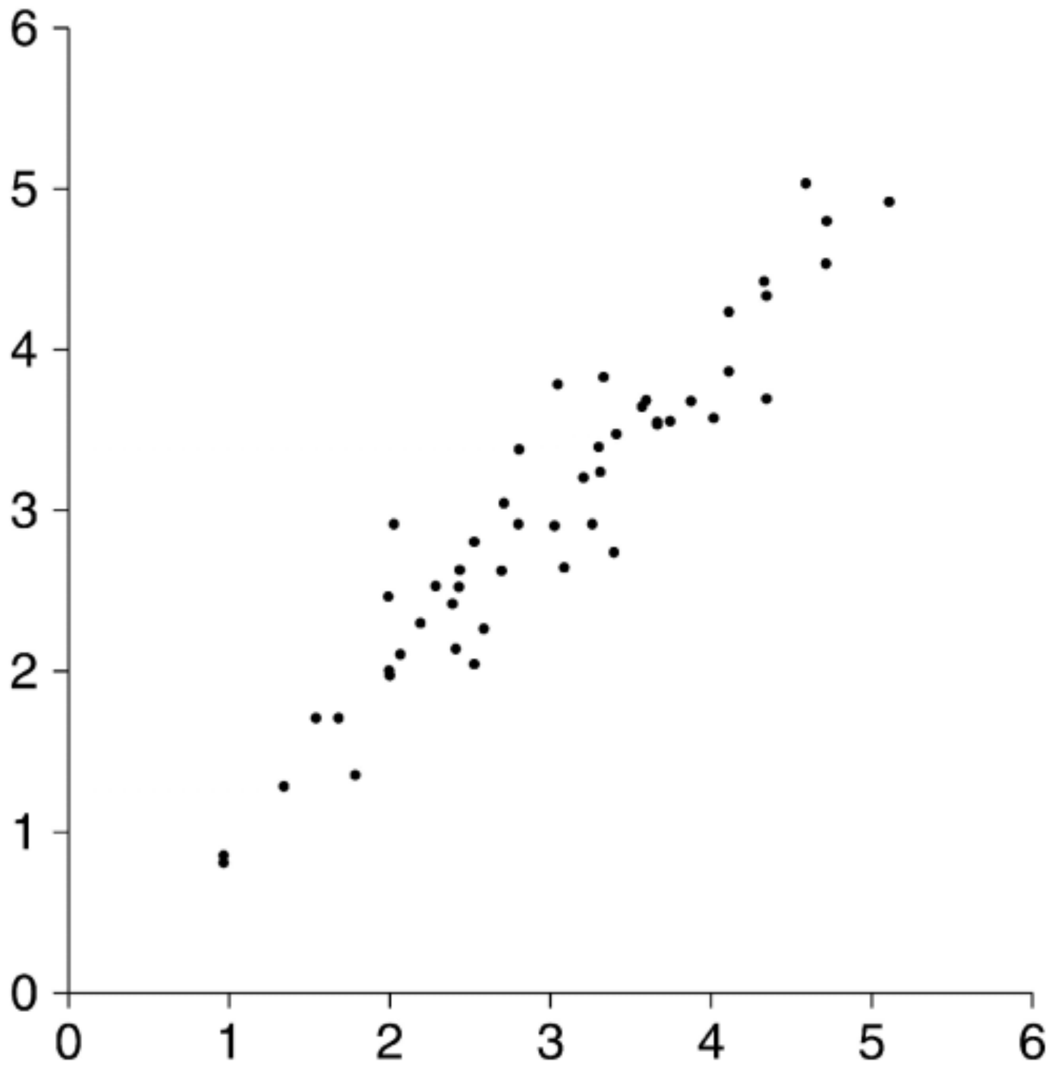
- *Positive:* y tends to get bigger when x is bigger.
- *Negative:* y tends to get smaller when x is bigger.
- *linear:* the points (x, y) in the scatterplot seem to cluster around a straight line.



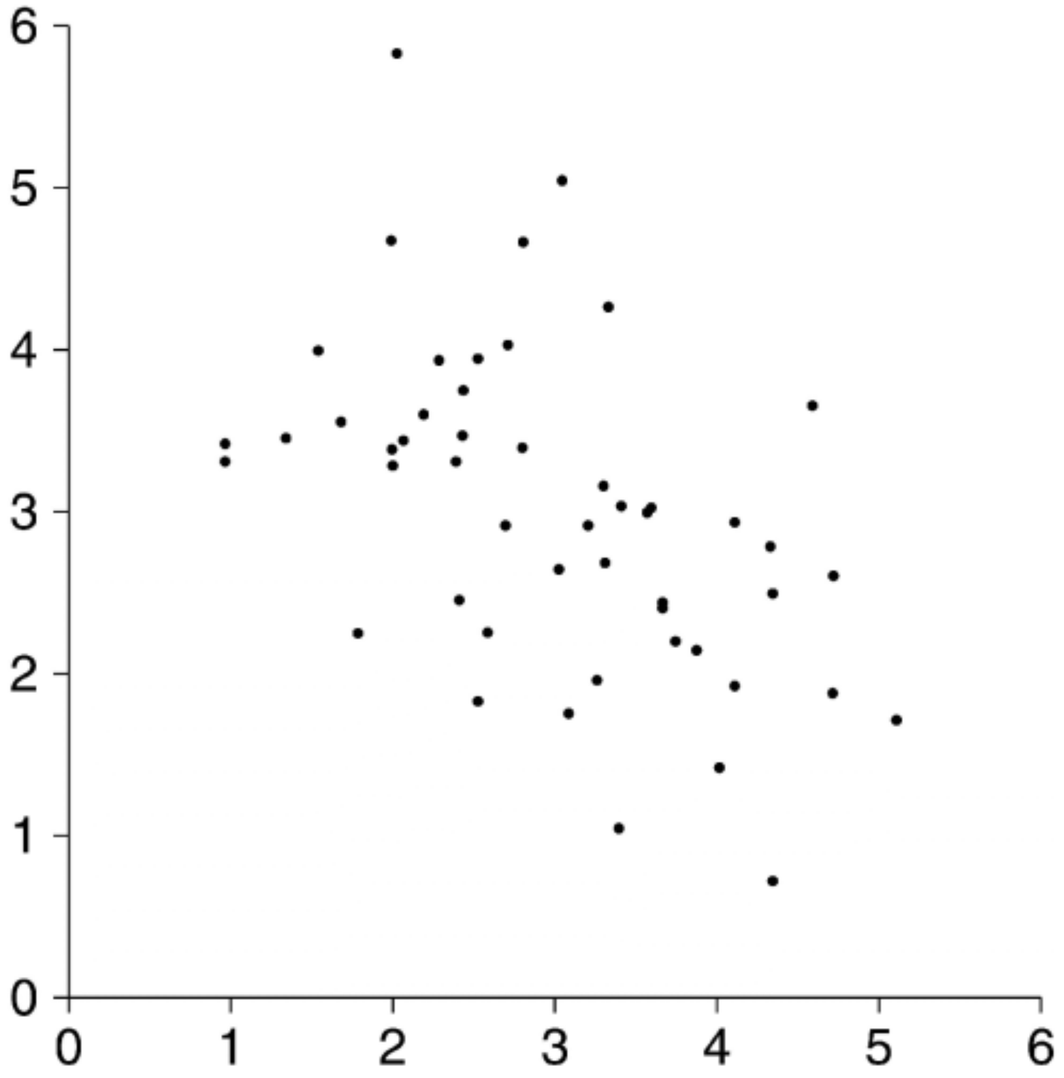
Weak positive association



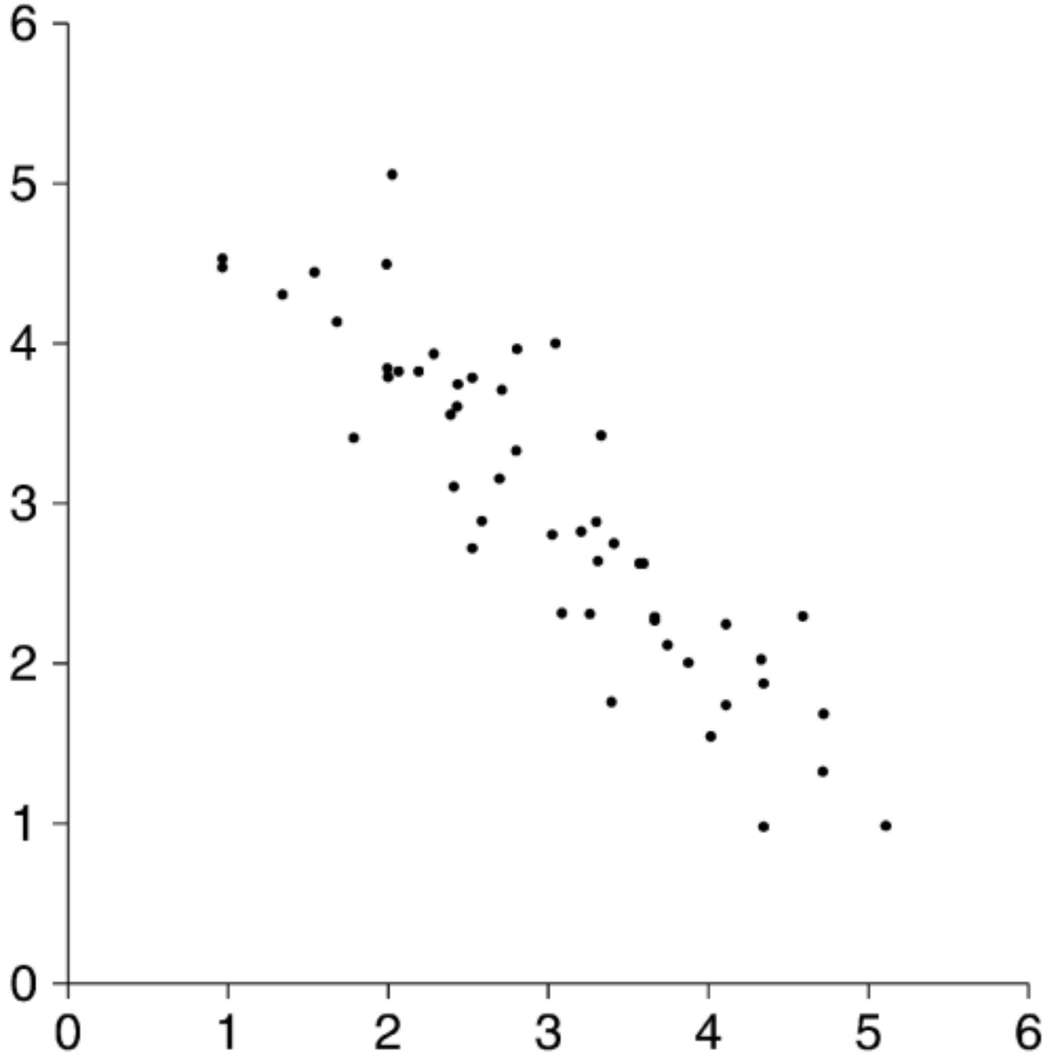
Stronger positive association



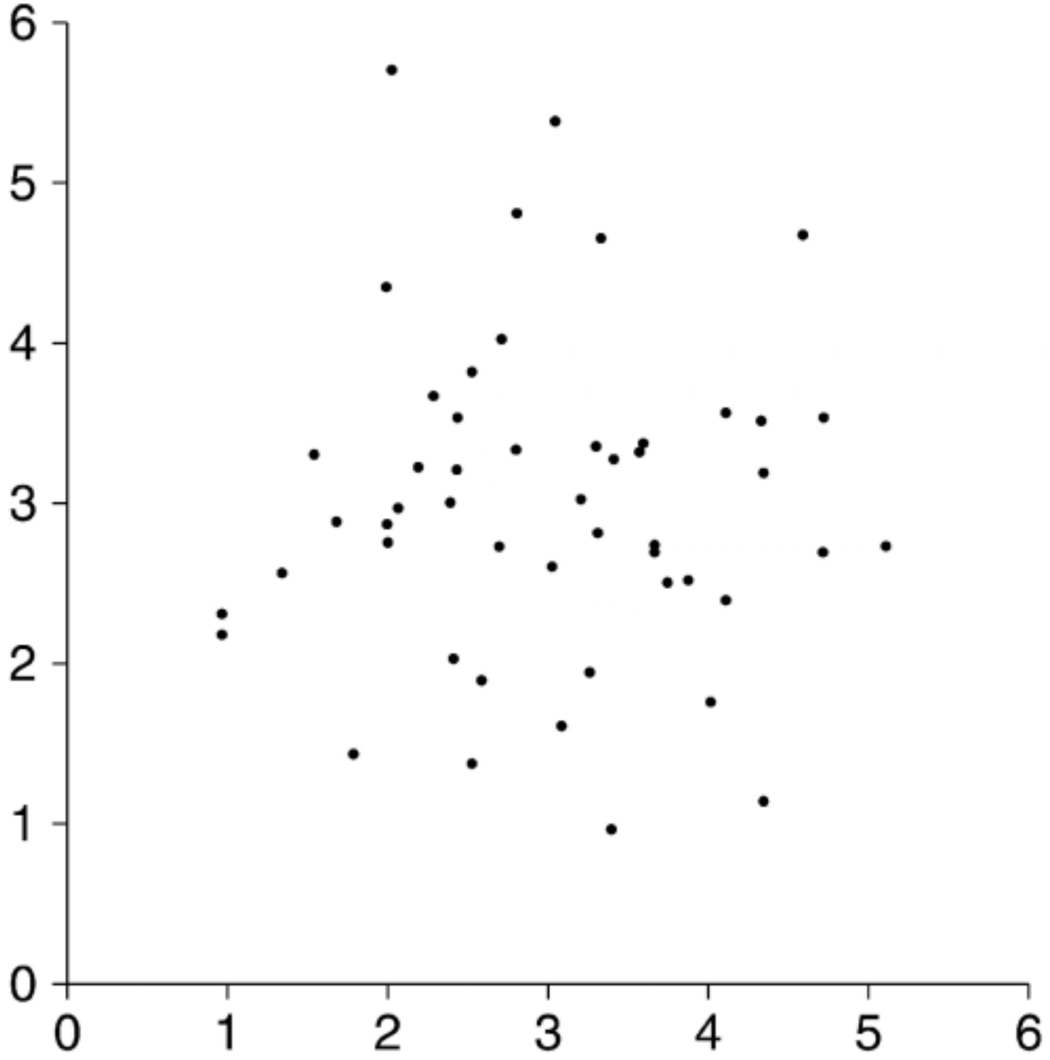
Very strong positive association



Weak negative association



Strong negative association



No obvious (linear) relationship

Comments:

(1) The relation between two variables is often complicated. In particular, many ‘dependent’ variables depend on more than just one ‘independent’ variable. This can make understanding the relation between just two variables more difficult.

(2) In many cases, a complicated relation between two variables can be *approximated* by a linear relation.

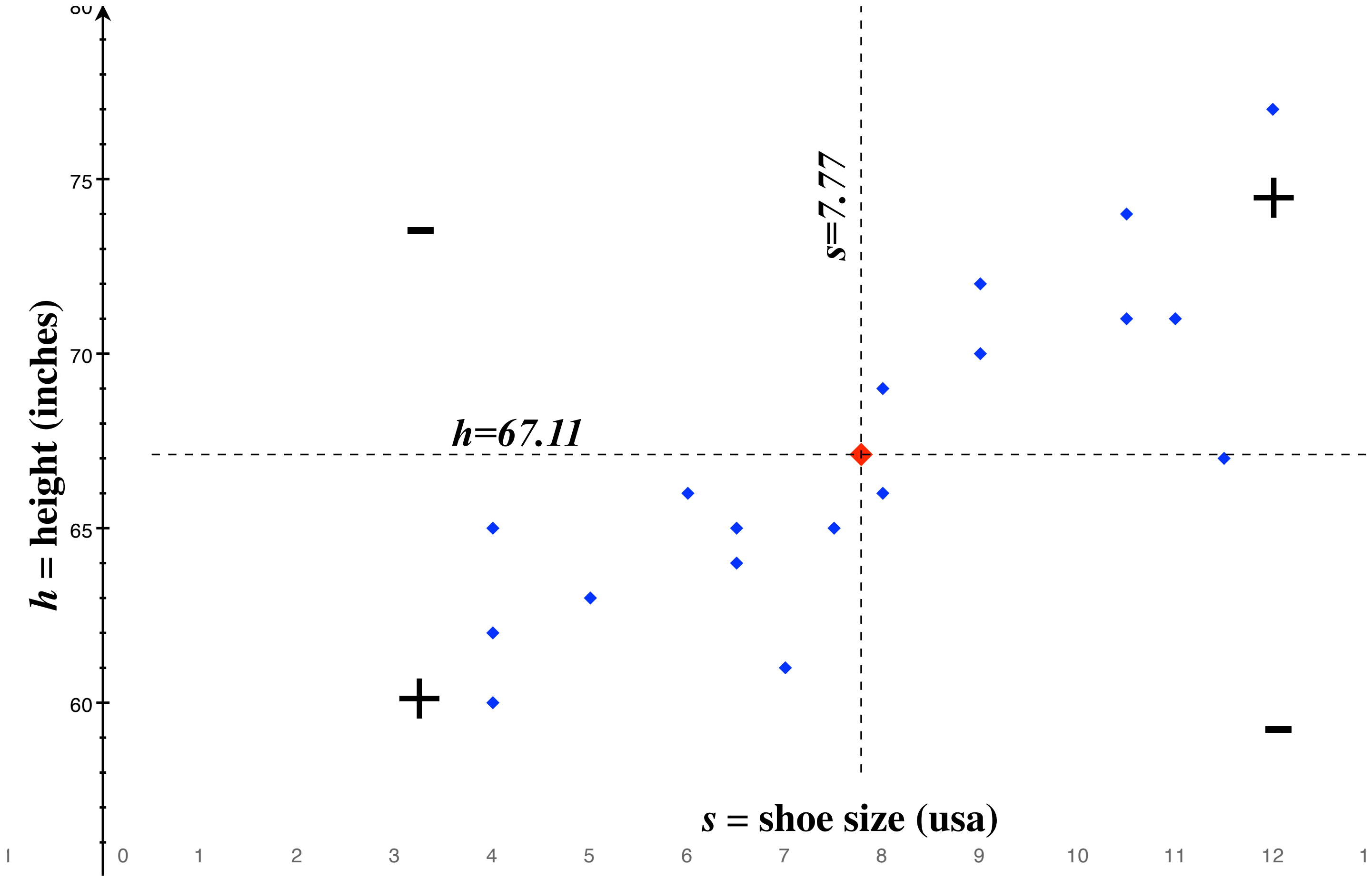
Want: a number that characterizes the nature and strength of the (linear) relation between two variables.

Observation:

(*) If the relation between x and y is *positive*, then above-average x-values will tend to be paired with above-average y-values and below-average x-values will tend to be paired with below-average y-values.

(*) If the relation between x and y is *negative*, then above-average x-values will tend to be paired with below-average y-values and below-average x-values will tend to be paired with above-average y-values.

The shoe size – height scatterplot with the point of averages $(\overline{s}, \overline{h})$ (the red diamond) and positive and negative quadrants (relative to the point of averages).



The correlation coefficient.

The ***correlation coefficient*** r_{xy} of a set of paired data,

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

is defined by

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

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

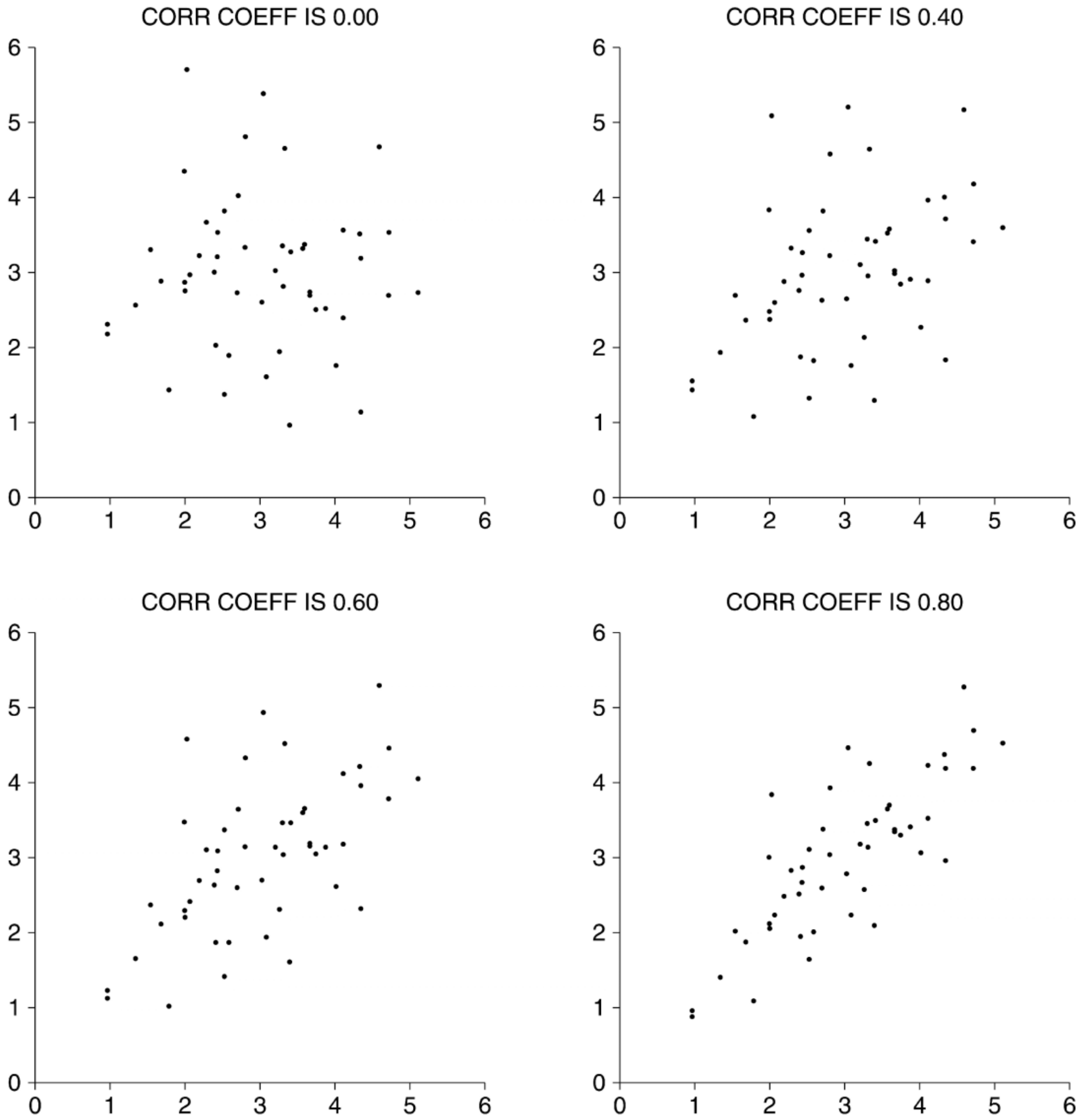
$$r_{xy} = \frac{1}{n} \sum z_{x_j} \cdot z_{y_j}.$$

Comment: Using products of the z-scores (instead of $(x_j - \overline{x})(y_j - \overline{y})$ by itself) makes the correlation coefficient insensitive to ***scale***.

In the height/shoe size example, the correlation is

$$\begin{aligned} r_{s,h} &= \frac{1}{18}\sum_{j=1}^{18}\left(\frac{s_j - 7.77}{2.58}\right)\cdot\left(\frac{h_j - 67.11}{4.54}\right) \\ &= \frac{1}{18}\left[\left(\frac{5-7.77}{2.58}\right)\cdot\left(\frac{63-67.11}{4.54}\right)+\cdots+\left(\frac{11-7.77}{2.58}\right)\cdot\left(\frac{71-67.11}{4.54}\right)\right] \\ &\approx 0.818. \end{aligned}$$

(*) The correlation is positive, as expected.

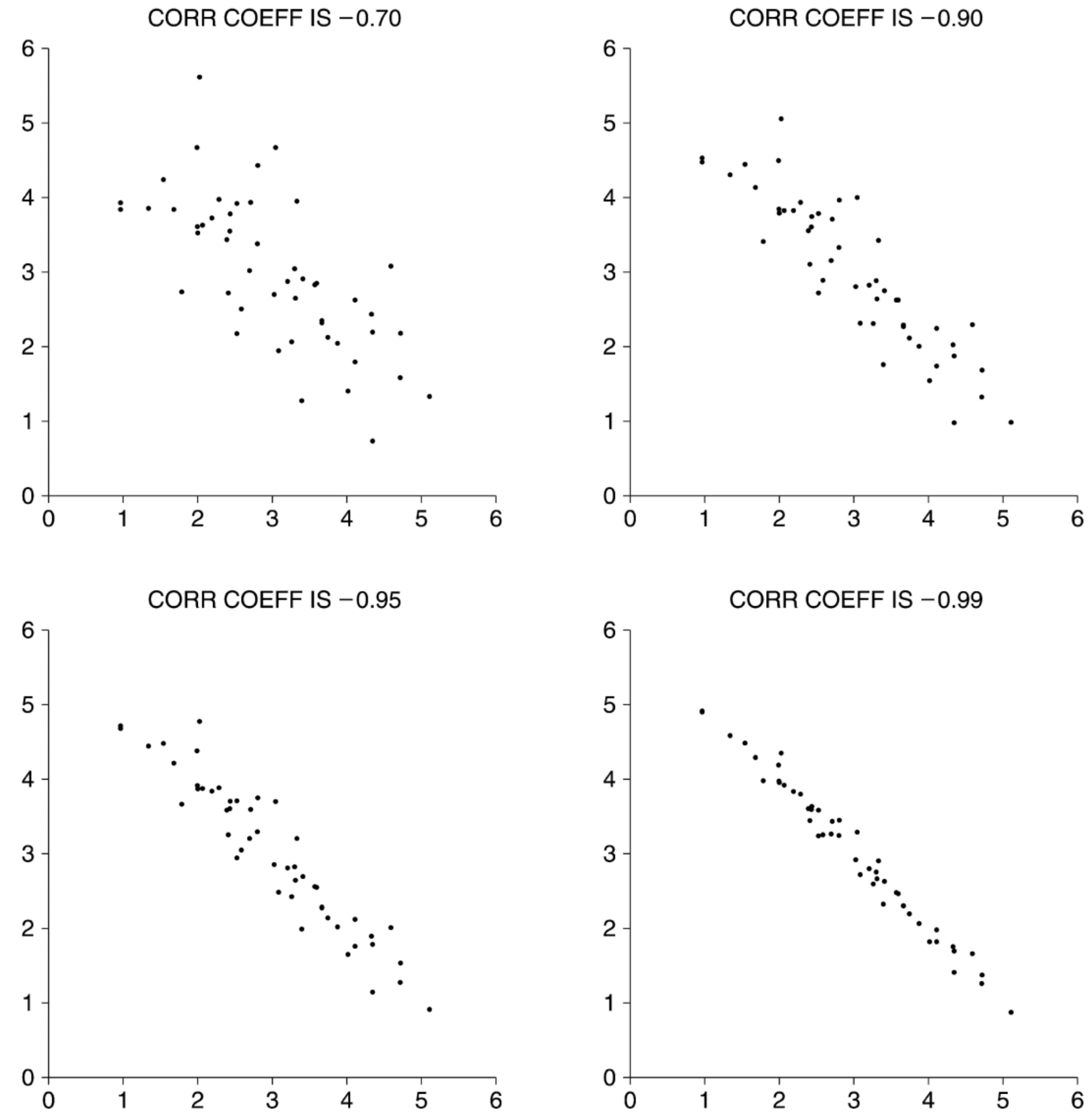


CORR COEFF IS 0.00

CORR COEFF IS 0.40

CORR COEFF IS 0.60

CORR COEFF IS 0.80



CORR COEFF IS −0.70

CORR COEFF IS −0.90

CORR COEFF IS −0.95

CORR COEFF IS −0.99

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.
- If $r_{xy} = \pm 1$, then the points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ all lie on the same straight line, and vice versa.

Question: *If there is strong correlation between the variables x and y ($|r|$ closer to 1) does this imply that there is a causal relation between the variables?*

Answer: ***Not*** by itself.

(*) The correlation coefficient is a measure of statistical (linear) ***association***. It does not *prove* ***causation***.

(*) In many cases where there is strong correlation, there are also significant confounding variables.

(*) The correlation coefficient *is sensitive to the data.* If the (sample) data is biased, the (sample) correlation coefficient may not be reliable.

Example.

☞ Shoe size and reading ability.

Many observational studies have noted a positive correlation between shoe size and reading ability.

Does having bigger feet make someone a better reader?

Probably ***not***. Older humans have both bigger feet and are better readers. Both shoe size and reading ability increase with age ***in children***.

Example.

☞ Education level and unemployment.

During the Great Depression (1929 - 1933), people with more education tended to be unemployed for shorter periods (on average).

Does education protect against unemployment?

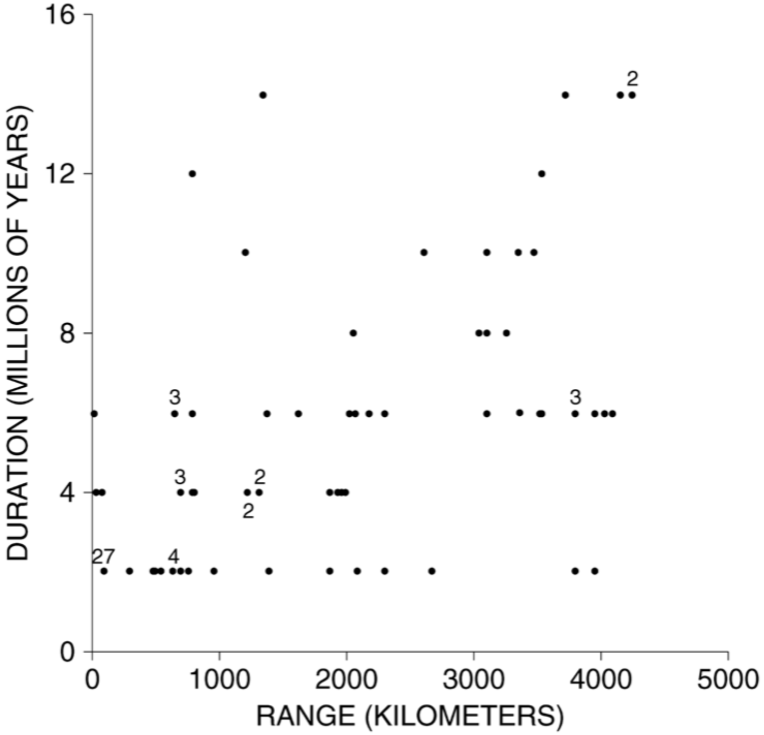
A little, but age is once again a confounding variable. Younger adults tended to have more education than older adults and employers tended to prefer hiring younger people.

Example.

➾ Does natural selection work at the level of species?

(Figure below from FFP, Chapter 9 – see the discussion there).

Figure 7. Duration of species in millions of years plotted against geographical range in kilometers, for 99 species of gastropods. Several species can be plotted at the same point; the number of such species is indicated next to the point.



Statistics, Fourth Edition
Copyright © 2007 W. W. Norton & Co., Inc.

Correlation: more observations.

(*) r_{xy} does not identify *nonlinear* relationships. E.g., you can have $r_{xy} \approx 0$, even though there is a very strong *nonlinear* relation between x and y.

(*) Because of the nature of the sample data, you can also have $|r_{xy}|$ relatively close to 1, even though the actual relation between x and y is nonlinear.

(*) Data with significant outliers are not well-described by the correlation coefficient.

(*) Correlation does not necessarily imply causation, and even when there is a causal relation, it should be interpreted carefully.

Example. In 2005, the correlation between age and years of education completed for women age 25 and above was $r \approx -0.2$.

Does this mean that women become less educated as they grow older?

Correlation: more observations.

(*) r_{xy} does not identify *nonlinear* relationships. E.g., you can have $r_{xy} \approx 0$, even though there is a very strong *nonlinear* relation between x and y.

(*) Because of the nature of the sample data, you can also have $|r_{xy}|$ relatively close to 1, even though the actual relation between x and y is nonlinear.

(*) Data with significant outliers are not well-described by the correlation coefficient.

(*) Correlation does not necessarily imply causation, and even when there is a causal relation, it should be interpreted carefully.

Example. The correlation between age and years of education completed for women age 25 and above in 2005 was $r \approx -0.2$.

Does this mean that women become less educated as they grow older?

No. The data is *cross-sectional.* Older women tend to have completed fewer years of school than younger women.

Ecological correlations

An *ecological correlation* is one that measures the correlation between class averages across several classes, instead of measuring the correlation between the variables at the level of individual observations.

☞ Ecological correlations tend to exaggerate the strength of the relationship because averaging reduces the variation in the data.

Examples.

(*) According to the CPS of 2005, for men age 25-64 in the U.S., the correlation between years of education and income was $r \approx 0.42$. On the other hand, when you compute the averages for income level and years of education for each of the 50 states and DC, the correlation for these 51 points is $r \approx 0.7$.

(*) Hypothetical extreme case: If you compute the correlation between averages across *two* classes, the ecological correlation will always be ± 1.