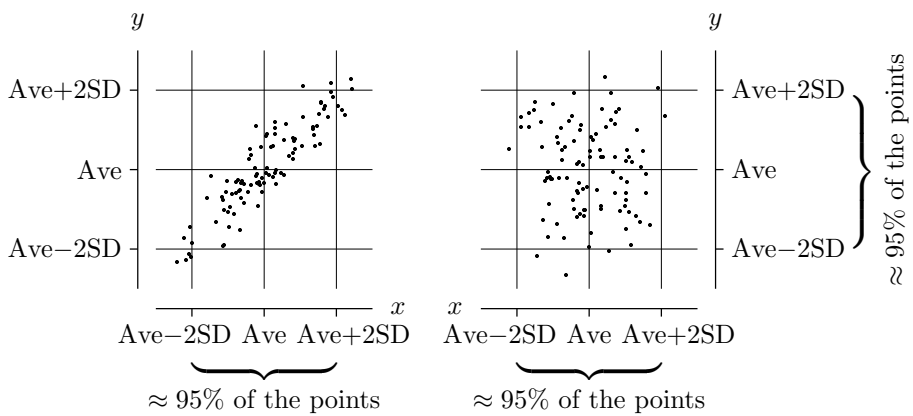
SUMMARIZING A SCATTER DIAGRAM

• A scatter diagram can be summarized by means of five statistics:

- The average and SD of the x-values.
- The average and SD of the y-values.
- The correlation coefficient r.

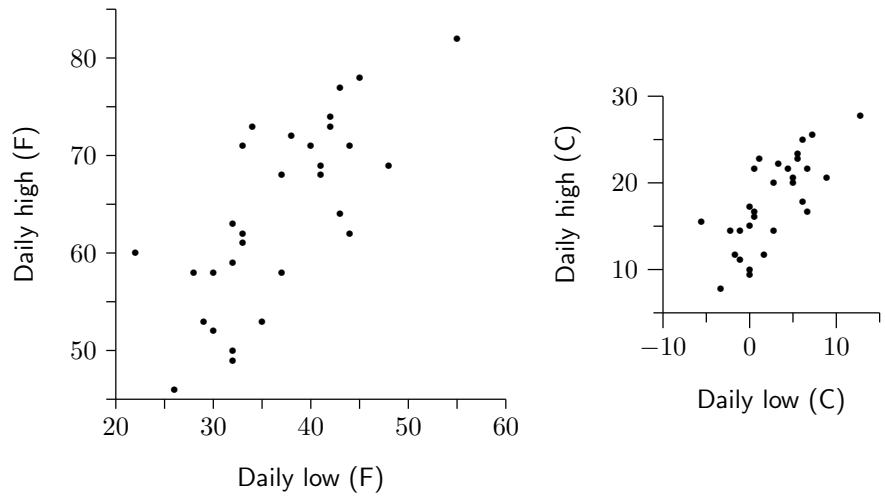
The averages and the SDs specify the location and spread of the cloud of points, horizontally and vertically. The correlation coefficient measures the amount of linear association, i.e., clustering about the SD line.



Strong association	Weak association
Tight clustering	Loose clustering
Knowing x helps in predicting y	Knowing x isn't much help in predicting y
Correlation near 1	Correlation near 0

SOME FEATURES OF THE CORRELATION COEFFICIENT

• Look at the following two scatter diagrams. They both show the daily low and high temperatures for Boulder CO for the month of April 1996. One uses degrees Fahrenheit, the other degrees Centigrade ($\mathrm{C} = \frac{5}{9}(\mathrm{F} - 32)$).

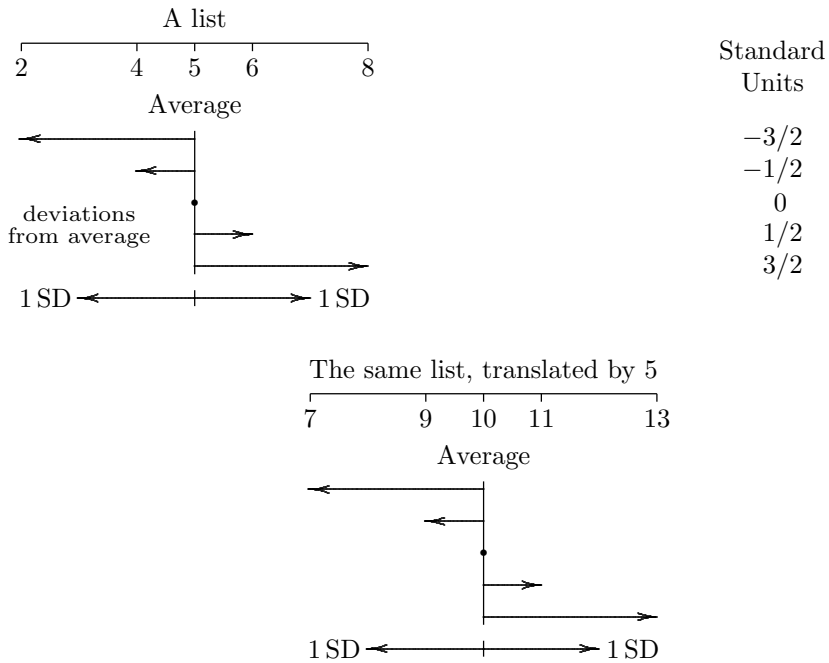


• Are the correlation coefficients the same, or different?

- They're the same; $r = 0.74$ for both diagrams.

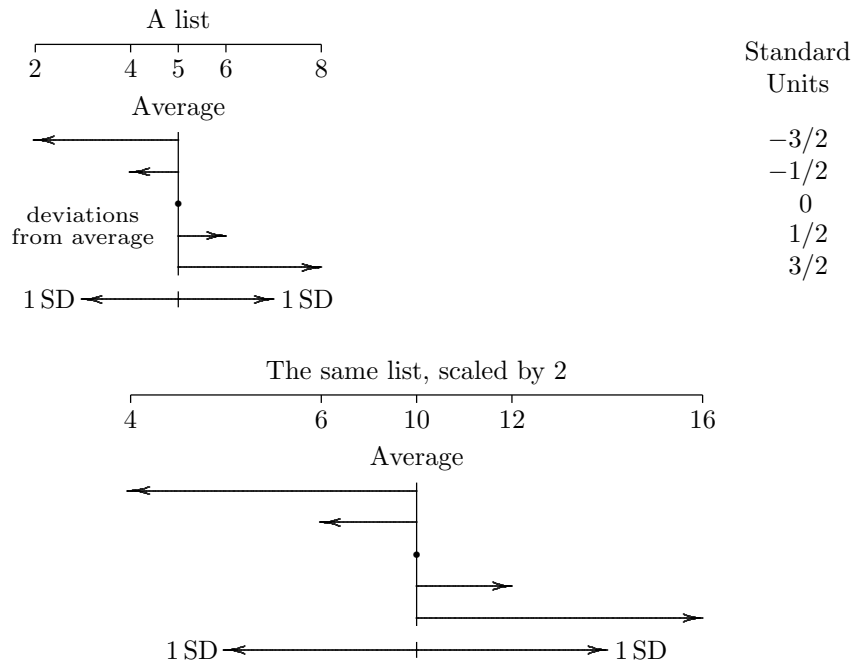
• To see why, we are going to consider the effect of *translation* (adding the same number to each value of a variable) and *scaling* (multiplying each value of a variable by the same positive number) on the correlation coefficient. First we need to consider the effect of translation and scaling on standard units.

- What effect does translation have on standard units?



- When you shift all the entries on a list by some number,
  - the average shifts by that number,
  - the deviations from average don't change,
  - the standard deviation doesn't change, and
  - the *standard units don't change*.
    - Standard units are deviations from the average, expressed in units of the SD.

- What effect does scaling have on standard units?



- When you multiply all the entries on a list by some positive number,
  - the average gets scaled by that number,
  - the deviations from average get scaled,
  - the standard deviation gets scaled, but
  - the *standard units don't change*.
    - The scale factor cancels out, when the deviations from the average are divided by the SD.

• What effect do translation and scaling have on the correlation coefficient?

• _____. The correlation r between two variables x and y is computed from the formula

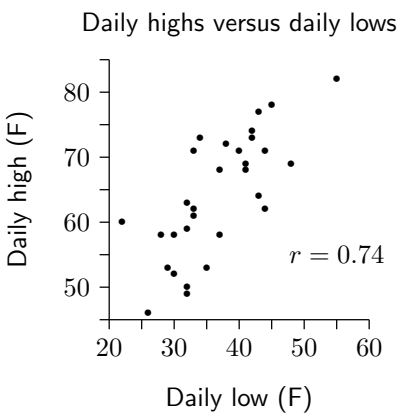
$$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units}).$$

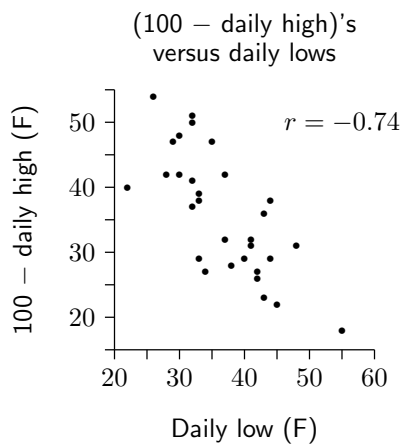
No matter how often you translate and scale x and y, the standard units won't change, and neither will r.

• The correlation between two sets of temperatures doesn't depend on whether the temperatures are measured in Fahrenheit or centigrade.

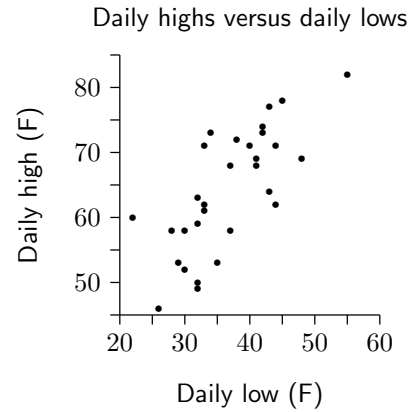
• Converting from Fahrenheit to centigrade involves a translation (subtract 32) followed by a scaling (multiply by 5/9).

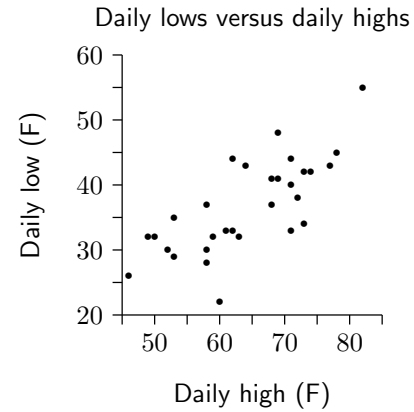
• But watch out — if you multiply all the values of one of the variables by a *negative* number, you'll *change the sign* of the correlation coefficient:





• Is the correlation between x and y the same as the correlation between y and x?





• _____ . The correlation between x and y involves averaging the products

$$(x \text{ in standard units}) \times (y \text{ in standard units}),$$

while the correlation between y and x involves averaging the products

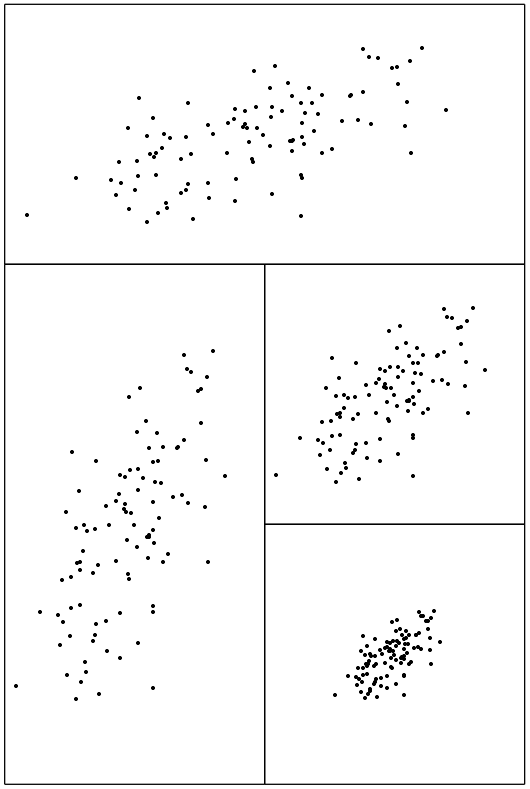
$$(y \text{ in standard units}) \times (x \text{ in standard units}).$$

But the order of the multiplications doesn't matter. E.g., if for some data point x is 3 SDs above average and y is 2 SDs below average, then

$$\begin{aligned}(x \text{ in standard units}) &\times (y \text{ in standard units})\\ &= \_\_\_\_\_ = -6 = -2 \times 3\\ &= (y \text{ in standard units}) \times (x \text{ in standard units}).\end{aligned}$$

• *Switching the axes has no effect on the correlation coefficient.*

• In which of the four diagrams below is the correlation coefficient the largest? The smallest?

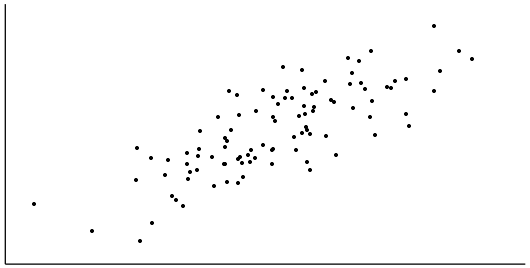


- Each diagram has $r = 0.6$. Only the SDs are different.
- Beware: SDs affect your perception of correlation.

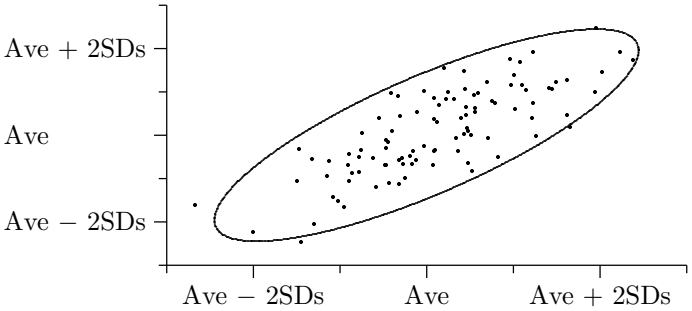
• The correlation coefficient measures clustering about the SD line, but only relative to the SDs of the variables involved.

GUESSING CORRELATION COEFFICIENTS BY EYE

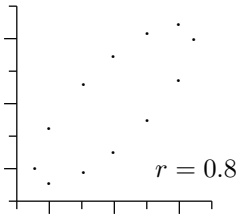
• How do you guess a correlation coefficient by eye, when the horizontal and vertical SDs have different lengths on the page?



• First lay out the standard units coordinate system and sketch the usual oval:



• Then resize the coordinate system to the same scale as used for the diagrams on pages 8–8 and 8–9 of the notes, and transfer the oval:



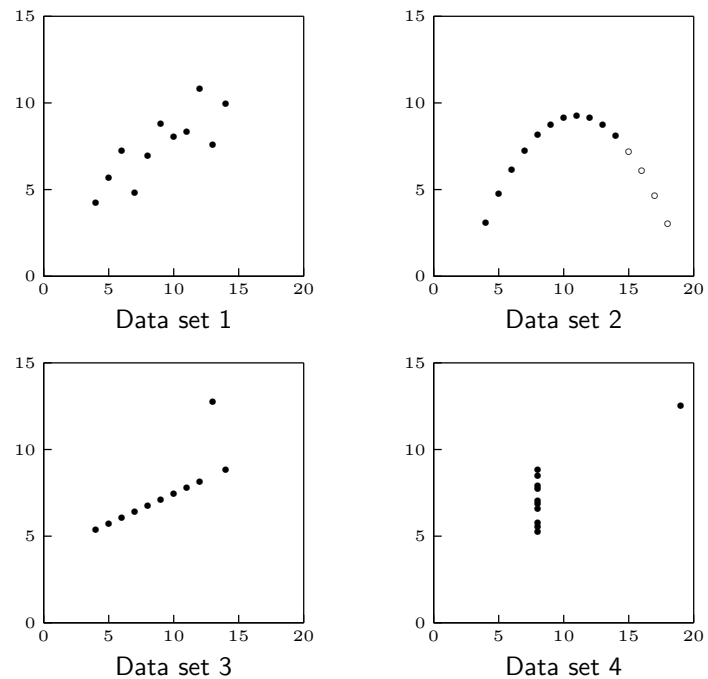
OUTLIERS AND NONLINEAR ASSOCIATION

- How are these 4 data sets alike?

	Data set 1		*Data set 2*		*Data set 3*		*Data set 4*	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.96	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.75	13	12.76	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.36	19	12.50
	12	0.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Ave	9	7.5	9	7.5	9	7.5	9	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

- How are the data sets different?

- The nature of the association between x and y is very different:

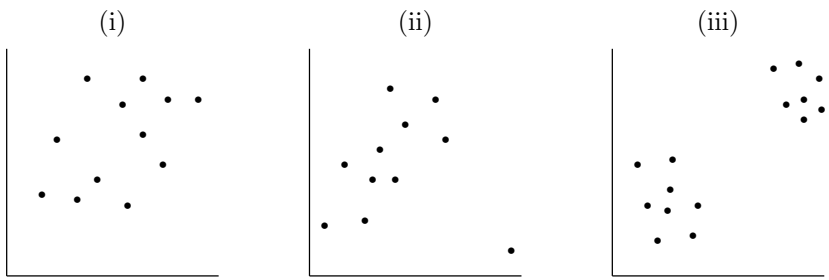


  - In the second data set, you could predict y exactly from x. But r is less than one† because the correlation coefficient only measures _____ association.
  - In the third data set, r would be _____ instead of 0.82 if the outlier were actually on the line.

- The correlation coefficient can be misleading in the presence of outliers or nonlinear association. Whenever possible, look at the scatter diagram to check for these problems.

† r drops to 0 when the four ○'s are included.

• Which of the following 3 scatter diagrams should be summarized by r?

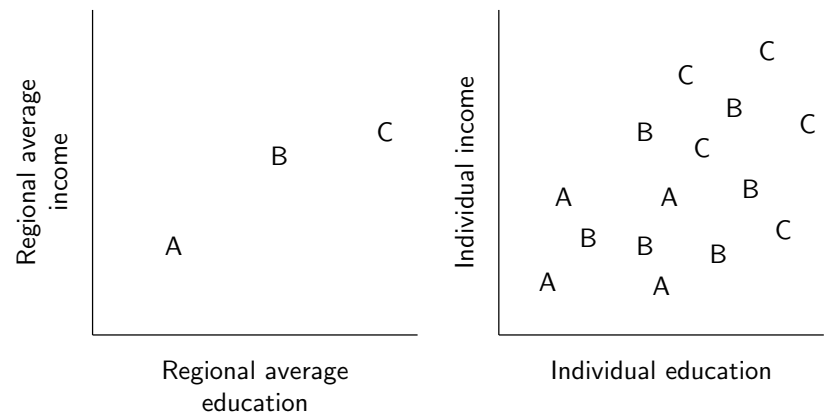


• Diagram (iii) has a curious property. Each of the two clusters exhibits a weak negative association. But the whole diagram shows a moderately strong positive association.

• A class of 15 students happens to include 5 basketball players. True or false, and explain: The relationship between heights and weights for this class can be summarized using r.

ECOLOGICAL CORRELATIONS

• For each of three geographical regions in the U.S., a political scientist computes the average income and average educational level for the men aged 25–64 living in those regions. The correlation between regional average education and regional average income is 0.9. Does this give a fair estimate of the strength of the association between education and income for the men being studied?



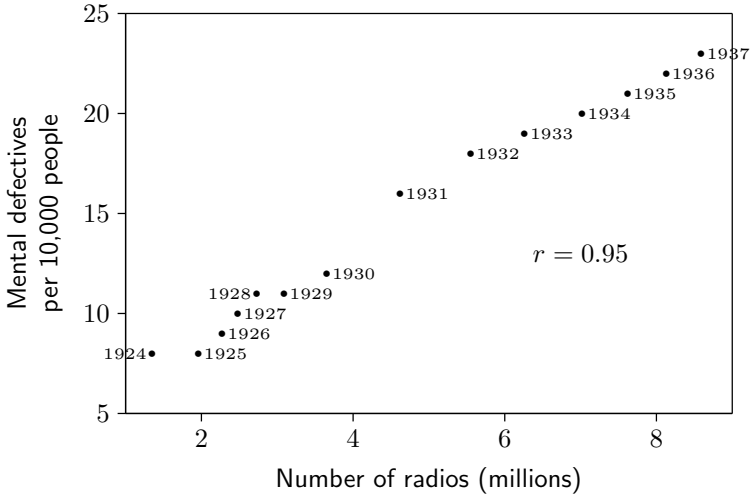
• No. Taking averages eliminates spread, and gives a misleading impression of tight clustering.

• *Ecological*† correlations, which are based on rates or averages, tend to overstate the strength of associations.

† ecology: n 1. ... 2. the branch of sociology concerned with the spacing of people and institutions and the resulting interdependency.

ASSOCIATION VERSUS CAUSATION

• The following scatter diagram shows the number of radios in the United Kingdom from 1924 to 1937 and the number of mental defectives per 10,000 people for the same years.



A social scientist explains this as follows: as more and more people gave up intellectual pursuits like reading for mindless listening to the radio, general atrophy of the brain set in and lead to increased mental disability. What do you say?

• Technological progress is a confounding factor. Over the years, there was a steady growth in the broadcasting industry and spread of the practice of listening to the radio. Concurrently, there was an increase in the interest in psychological ailments and the facilities for their treatment.

• Children's shoe size and reading skill are strongly correlated.

- Does this mean that we should stretch children's feet to get them to read better?
- Does having longer feet improve your reading skill?
- What does it mean?

• Many studies have found an association between cigarette smoking and heart disease. One study found an association between coffee drinking and heart disease. Should you conclude that coffee drinking causes heart disease? Or can you explain the association between coffee drinking and heart disease in some other way?

• Correlation measures association. But association does not necessarily show causation. It may only show that both variables are simultaneously influenced by some third variable.

SUMMARY

- The correlation coefficient is a pure number, without units. It is not affected by:
  - interchanging the two variables;
  - adding the same number to all the values of one variable;
  - multiplying all the values of one variable by the same positive number.
- The correlation coefficient measures clustering about the SD line, but only relative to the SDs of the variables involved.
- The correlation coefficient can be misleading in the presence of outliers or nonlinear association. Whenever possible, look at the scatter diagram to check for these problems.
- Ecological correlations, which are based on rates or averages, tend to overstate the strength of associations.
- Correlation measures association. But association does not necessarily show causation. It may only show that both variables are simultaneously influenced by some third variable.