Stats1 Chapter 4: Correlation

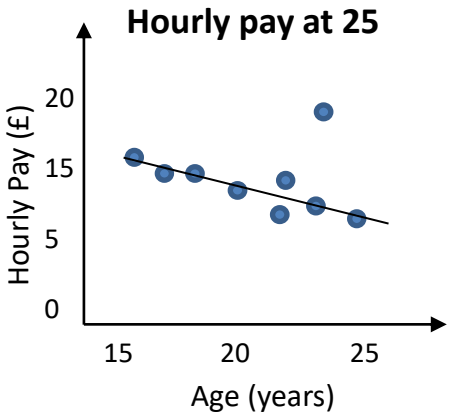
Correlated Variables

This Chapter Overview

Previously we have only considered one variable at a time. When we introduce a second variable (e.g. height with age), **we might want to consider the relationship between them**.

This is a short chapter!

"Describe the type of correlation."



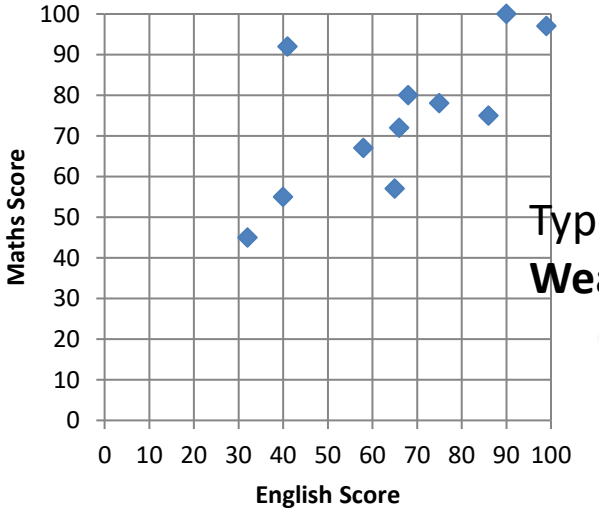
"The daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded. The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$.

(a) Given an interpretation of the value of the gradient of this regression line.

(b) Justify the use of a linear regression line in this instance."

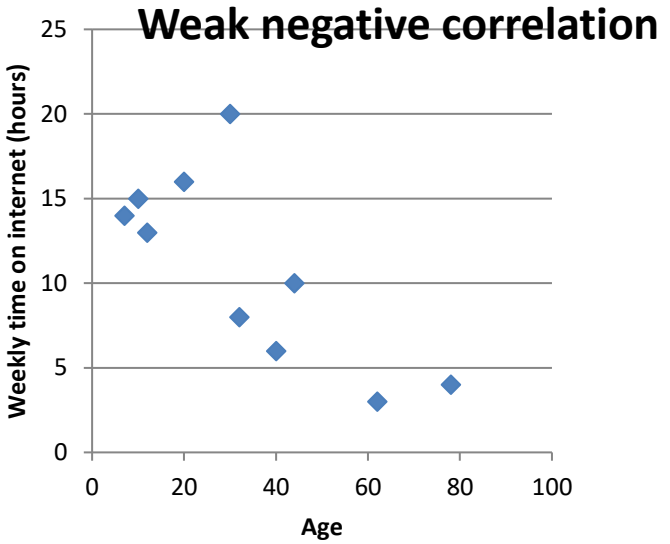
Recap of correlation

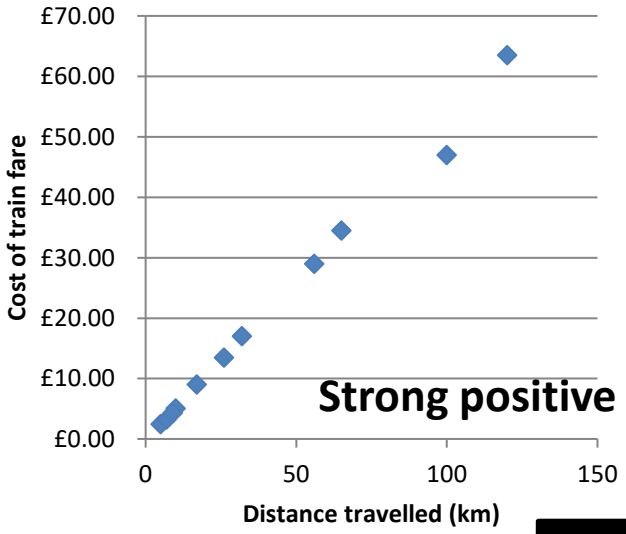
Correlation gives the **strength of the relationship** (and the type of relationship) between two variables. Data with two variables is known as **bivariate data**.

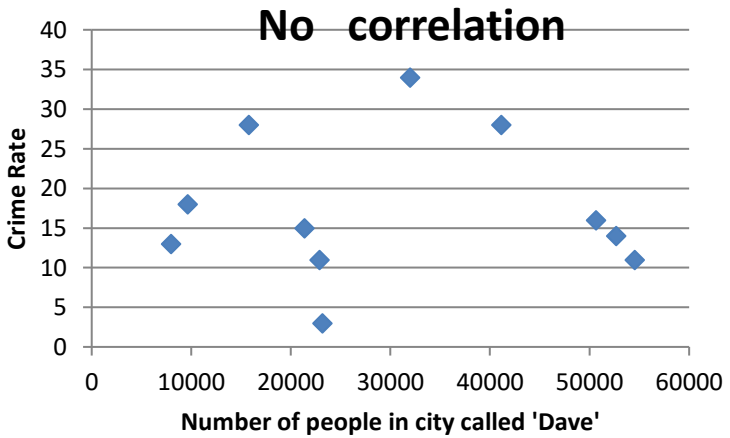


Type of correlation:
Weak positive correlation

strength | type





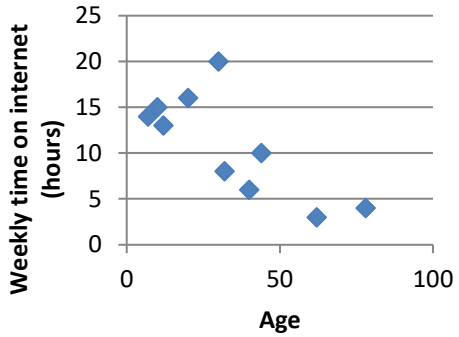


The vertical-axis variable usually **depends** on the horizontal-axis value. For this reason distance would be the **independent/explanatory variable** and cost the **dependent/response variable**.

Important correlation concepts

Important Point 1

To **interpret** the correlation between two variables is to give a worded description in the context of the problem.



a) State the correlation shown.
b) Describe/interpret the relationship between age and weekly time on the internet.

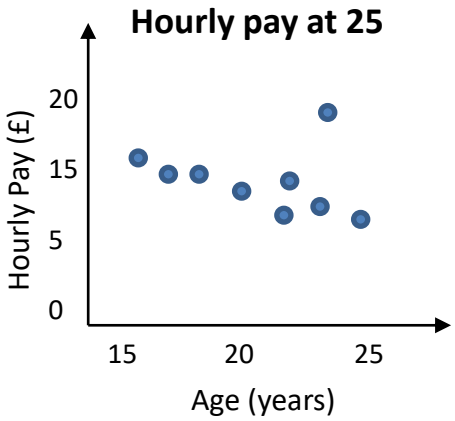
a) ?
b) ?

Important Point 2

[Textbook] Two variables have a **causal relationship** if a change in one variable directly causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

Hideko was interested to see if there was a relationship between what people earn and the age which they left education or training. She says her data supports the conclusion that more education causes people to earn a lower hourly rate of pay. Give one reason why Hideko's conclusion might not be valid.

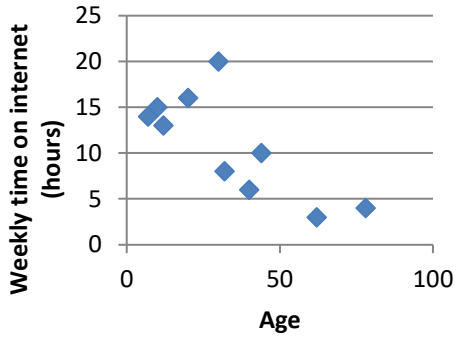
?



Important correlation concepts

Important Point 1

To **interpret** the correlation between two variables is to give a worded description in the context of the problem.



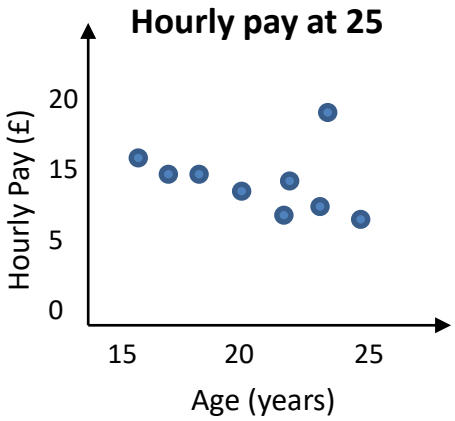
a) State the correlation shown.
b) Describe/interpret the relationship between age and weekly time on the internet.

a) Negative correlation.
b) As age increases, the weekly time on the internet tends to decrease.

Important Point 2

[Textbook] Two variables have a **causal relationship** if a change in one variable directly causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

Hideko was interested to see if there was a relationship between what people earn and the age which they left education or training. She says her data supports the conclusion that more education causes people to earn a lower hourly rate of pay. Give one reason why Hideko's conclusion might not be valid.



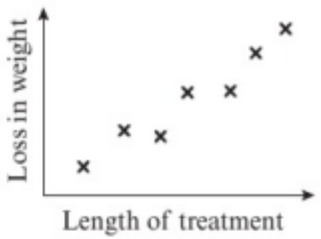
"Respondents who left education later would have significantly less work experience than those who left education earlier. This could be the cause of the reduced income shown in her results."

Exercise 4.1

Pearson Statistics/Mechanics Year 1/AS
Pages 26-27

Homework Exercise

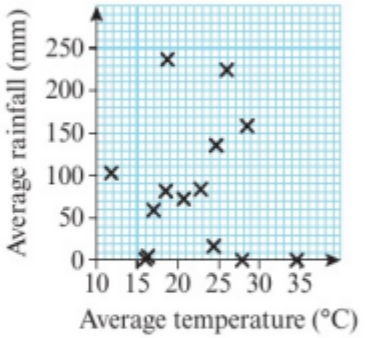
1 Some research was done into the effectiveness of a weight-reducing drug. Seven people recorded their weight loss and this was compared with the length of time for which they had been treated. A scatter diagram was drawn to represent this data.



a Describe the type of correlation shown by the scatter diagram.

b Interpret the correlation in context.

2 The average temperature and rainfall were collected for a number of cities around the world.
The scatter diagram shows this information.



a Describe the correlation between average temperature and average rainfall.

b Comment on the claim that hotter cities have less rainfall.

3 Eight students were asked to estimate the mass of a bag of sweets in grams. First they were asked to estimate the mass without touching the bag and then they were told to pick the bag up and estimate the mass again. The results are shown in the table below.

Student	A	B	C	D	E	F	G	H
Estimate of mass not touching bag (g)	25	18	32	27	21	35	28	30
Estimate of mass holding bag (g)	16	11	20	17	15	26	22	20

a Draw a scatter diagram to represent this data.

b Describe and interpret the correlation between the two variables.

Homework Exercise

4 Donal was interested to see whether there was a relationship between the value of a house and the speed of its internet connection, as measured by the time taken to download a 100 megabyte file. The table shows his results.

Time taken (s)	5.2	5.5	5.8	6.0	6.8	8.3	9.3	13	13.6	16.0
House value (£1000s)	300	310	270	200	230	205	208	235	175	180

a Draw a scatter diagram to represent this data.

b Describe the type of correlation shown.

Donal says that his data shows that a slow internet connection reduces the value of a house.

c Give one reason why Donal's conclusion may not be valid.

5 The table shows the daily total rainfall, r mm, and daily total hours of sunshine, s, in Leuchars for a random sample of 11 days in August 1987, from the large data set.

r	0	6.8	0.9	4.8	0	21.7	1.7	4.9	0.1	2.2	0.1
s	8.4	4.9	10.2	4.5	3.3	3.9	5.4	1.8	9.7	1	4.6

The median and quartiles for the rainfall data are: $Q_1 = 0.1$ $Q_2 = 1.7$ $Q_3 = 4.85$

An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

a Show that $r = 21.7$ is an outlier. **(1 mark)**

b Give a reason why you might:

i include **ii** exclude this day's readings. **(2 marks)**

c Exclude this day's readings and draw a scatter diagram to represent the data for the remaining ten days. **(3 marks)**

d Describe the correlation between rainfall and hours of sunshine. **(1 mark)**

e Do you think there is a causal relationship between the amount of rain and the hours of sunshine on a particular day? Explain your reasoning. **(1 mark)**

Homework Answers

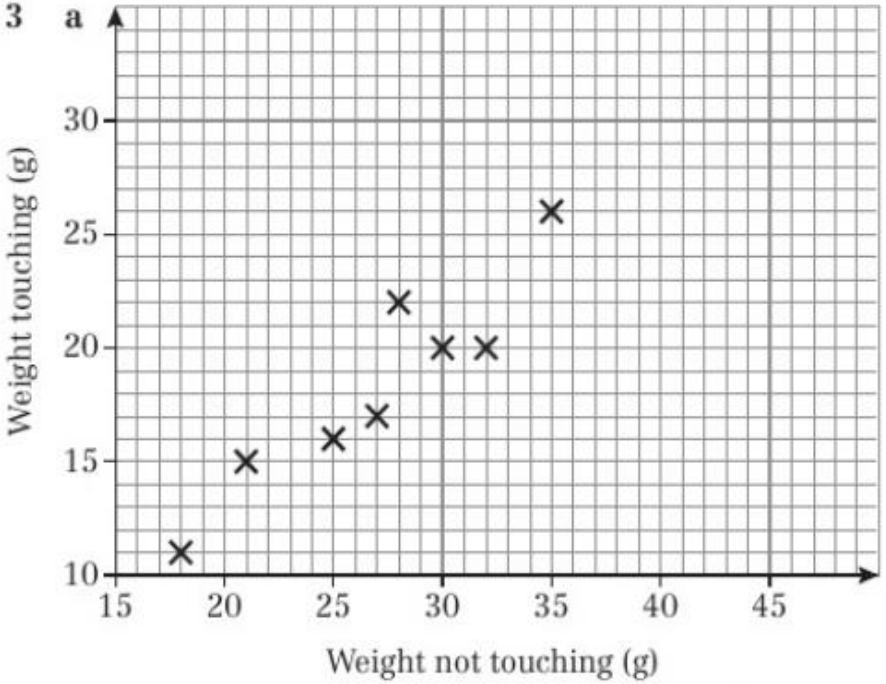
1 **a** Positive correlation.

b The longer the treatment, the greater the loss of weight.

2 **a** No correlation.

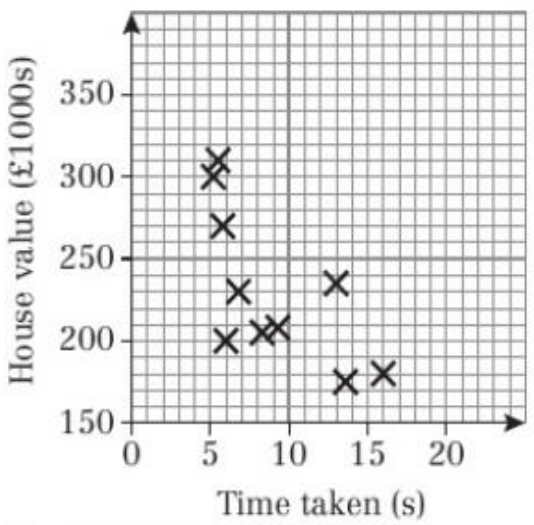
b The scatter graph does not support the statement that hotter cities have less rainfall.

3 **a**



b There is positive correlation. If a student guessed a greater weight before touching the bag, they were more likely to guess a greater weight after touching it.

4 **a**



b Weak negative correlation.

c For example, there may be a third variable that influences both house value and internet connection, such as distance from built up areas.

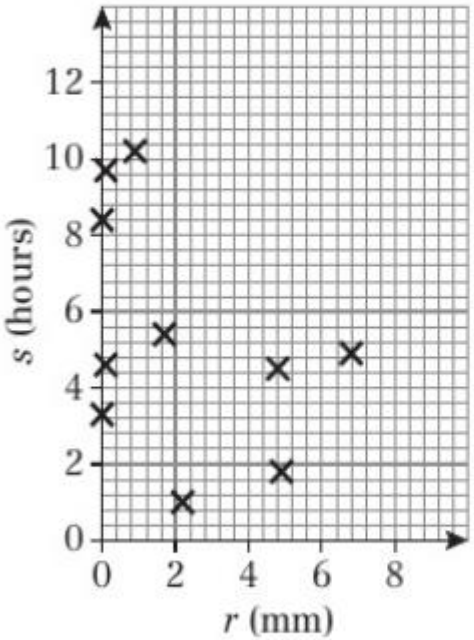
Homework Answers

5 **a** $Q_3 + 1.5 \times IQR = 4.85 + 7.125 = 11.975$
$21.7 > 11.975$, therefore is an outlier.

b **i** There is no reason to believe that the data collected by the Met Office is incorrect.

ii 21.7 is an outlier so may not be representative of the typical rainfall.

c



d Weak negative correlation.

e For example, there could be a causal relationship as days with more rainfall will have more clouds, and therefore less sunshine.