Chapter 4 - Statistics

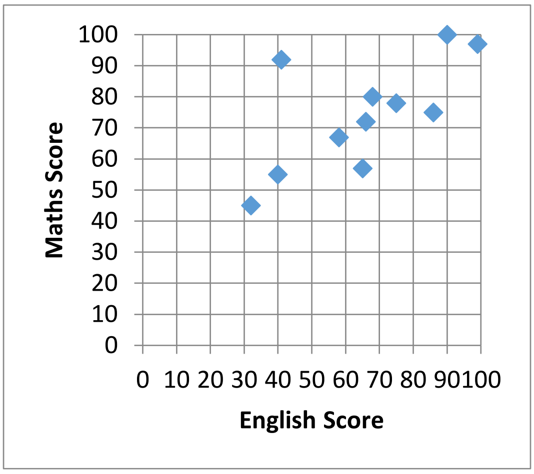
Correlation

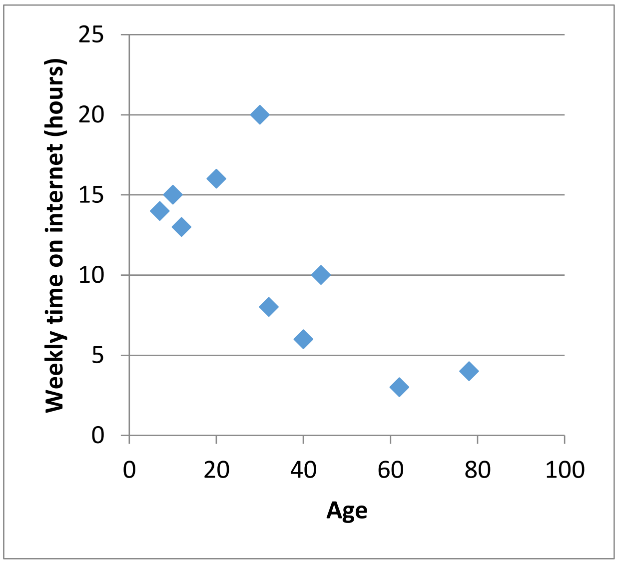
Chapter Overview

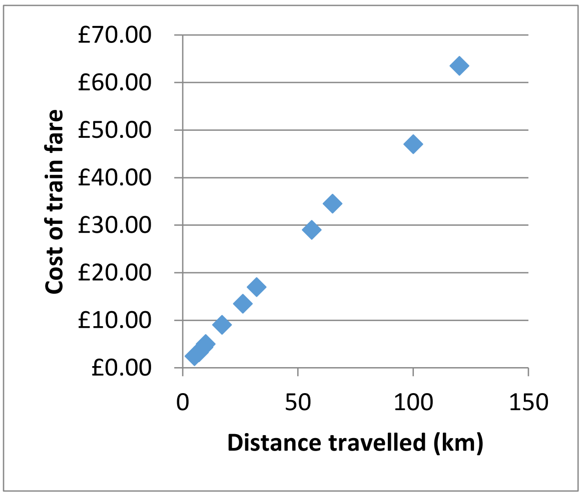
1. Draw and interpret scatter diagrams
2. Interpret correlation
3. Interpret the coefficients of a regression line equation for bivariate data
4. Understand when you can use a regression line to make predications

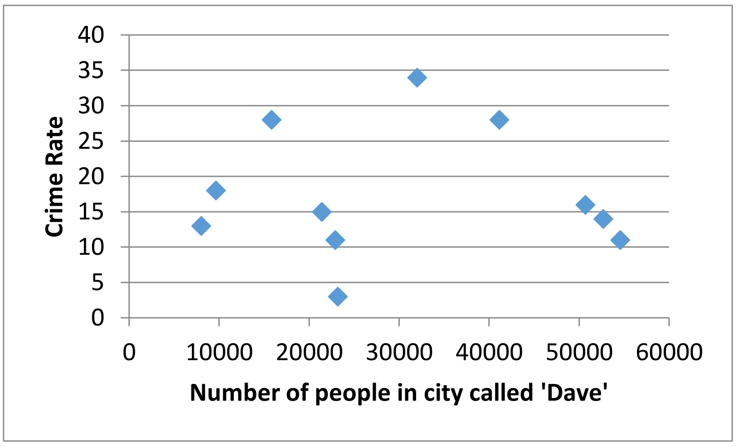
Topics	What students need to learn:		
	Content		Guidance
2 **Data presentation and interpretation** *continued*	2.2	**Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population (calculations involving regression lines are excluded).**	**Students should be familiar with the terms explanatory (independent) and response (dependent) variables.** Use to make predictions within the range of values of the explanatory variable and the dangers of extrapolation. Derivations will not be required. Variables other than x and y may be used. **Use of interpolation and the dangers of extrapolation. Variables other than x and y may be used.** Change of variable may be required, e.g. using knowledge of logarithms to reduce a relationship of the form $y = ax^n$ or $y = kb^x$ into linear form to estimate a and n or k and b.
		Understand informal interpretation of correlation. **Understand that correlation does not imply causation.**	**Use of terms such as positive, negative, zero, strong and weak are expected.**

Recap on Correlation





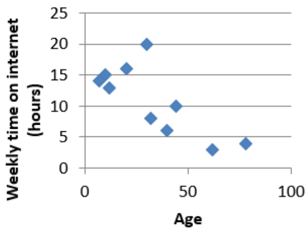




Important Correlation Concepts

Important Point 1

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

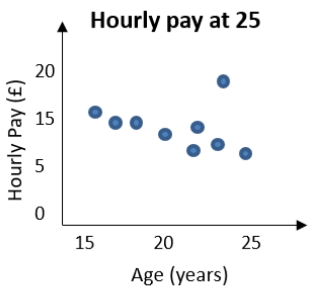


a) State the correlation shown.
b) Describe/interpret the relationship between age and weekly time on the internet.

Important Point 2

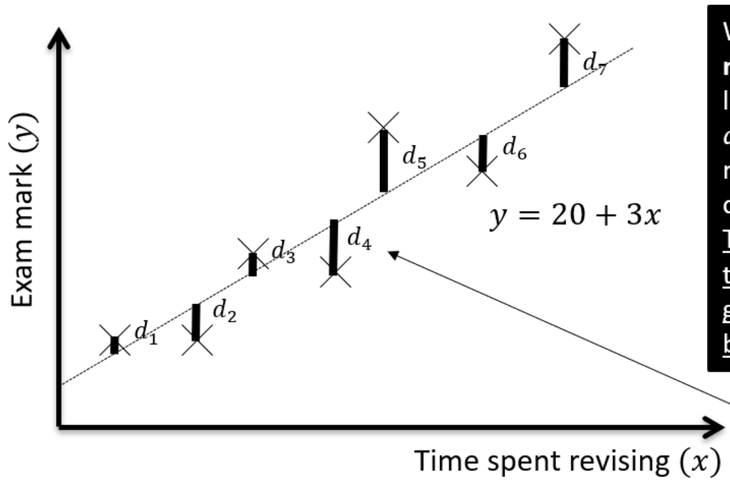
[Textbook] Two variables have a **causal relationship** if a change in one variable directly causes a change in the other. Just because two variables show correlation it does not necessarily mean that they have a causal relationship.

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.



Textbook Exercise 4A Pg 61

What is Regression?



What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set a and b such that the resulting y value matches the actual exam marks as closely as possible.
The 'regression' bit is the act of setting the parameters of our model (here the gradient and y-intercept of the line of best fit) to best explain the data.

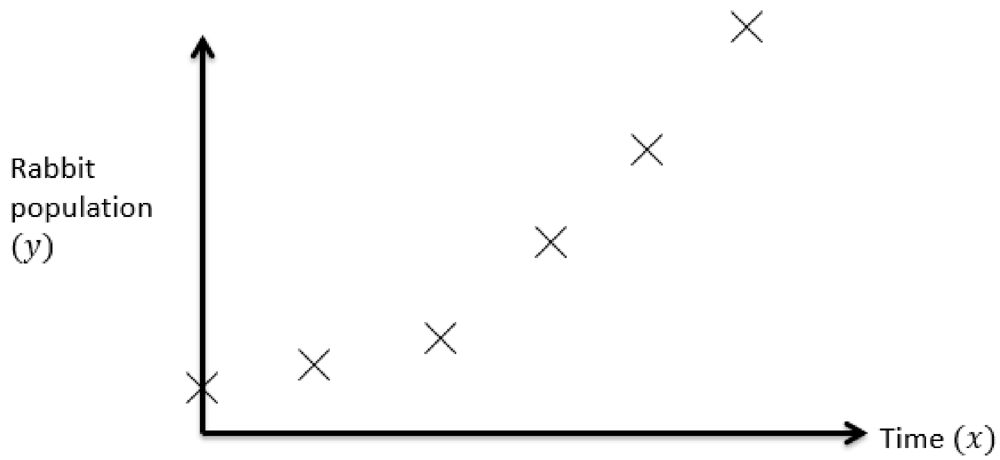
One type of line of best fit is the **least squares regression line**. This minimises the sum of the square of these 'errors', i.e.

$$d_1^2 + d_2^2 + \cdots = \Sigma d_i^2$$

Part of the reason we square these errors is so that each distance is treated as a positive value.

Unlike in the old S1, you are no longer required to work out the equation of the least squares regression line yourself; you will be given the equation.

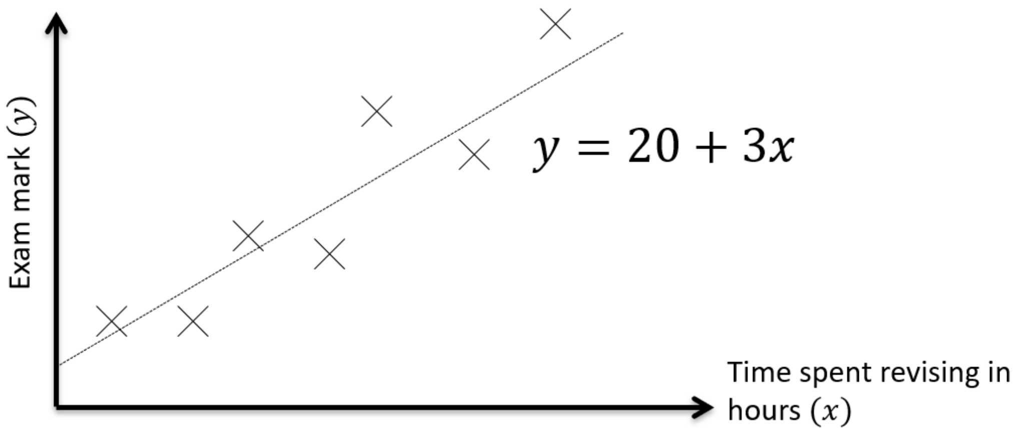
I record people's exam marks as well as the time they spent revising. I want to predict how well someone will do based on the time they spent revising.



In this chapter we only cover **linear regression**, where our chosen model is a straight line.

But in general we could use any model that might best explain the data. Population tends to grow exponentially rather than linearly, so we might make our model $y = a \times b^x$ and then try to use regression to work out the best a and b to use. **You will do exponential regression in Chapter 14 of Pure Year 1.**

Interpreting a and b.



How do we interpret the gradient of 3?

How do we interpret the y-intercept of 20?

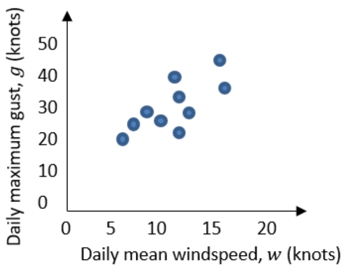
Example

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

© Met Office

The data was plotted on a scatter diagram.



(a) Describe the correlation between daily mean windspeed and daily maximum gust.

The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

(b) Give an interpretation of the value of the gradient of this regression line.

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

a

b

c

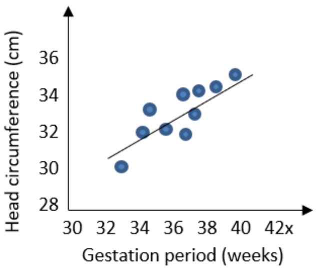
Interpolating and Extrapolating

Interpolating =

Extrapolating =

Example

[Textbook] The head circumference, y cm, and gestation period, x weeks, for a random sample of eight newborn babies at a clinic are recorded.
The scatter graph shows the results.
The equation of the regression line of y on x is $y = 8.91 + 0.624x$. The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.
(a) Comment on the reliability of these estimates.
A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.
(b) Explain why the regression equation given above is not suitable for this estimate.



Textbook Exercise 4B Pg 65

<u>Using your Classwiz</u>

Verify the regression line of g on w for the data above has the equation g=7.23 + 1.82w.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The Large Data Set

Locations

5 UK weather stations

3 overseas

Time Periods

May – October 1987 (6 months)
May – October 2015 (6 months)

Seasons

May/June are the end of spring
July-Sept is summer
October is autumn

Perth (Australia) is in the southern hemisphere, so July-Sept is winter

UK Great Storm

The night of 15-16th October 1987
Gusts up to 100 knots recorded

Florida hurricanes

12 October 1987 Hurricane Floyd
1-2 October 2015 Hurricane Joaquin

Variables Recorded

Daily Maximum Temperature
°C

Daily Total Rainfall
mm

Daily Total Sunshine
hours

Daily Maximum Relative Humidity
%; mist and fog if > 95%

Daily Mean Windspeed; Daily Maximum Gust
knots (1kn = 1.15mph)
and Beaufort scale

Daily Mean Wind Direction; Daily Maximum Gust Direction
bearing (°)
and cardinal direction

Cloud Cover
oktas (eights); 0 – 8

Visibility
Dm (decametres)
1 Dm =10m

Pressure
hPa (hectoPascal)

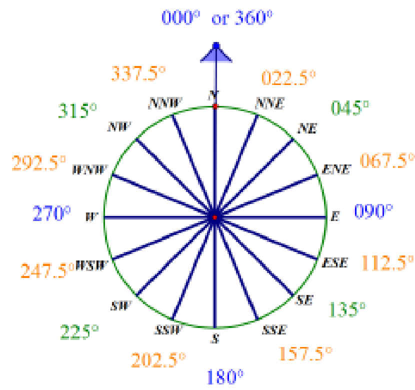
n/a
reading not available

tr (trace)
rainfall < 0.05mm

Beaufort Scale

Discrete, scale of 13 values:
0 (calm, < 1kn)
12 (hurricane, 64kn+)

Cardinal Directions



Oktas

Eighths of the sky covered by cloud
Discrete, scale of 9 values:
0 (clear sky)
8 (completely overcast)

Sources
Maps: *Pearson*
Compass: *mathsmutt.co.uk*