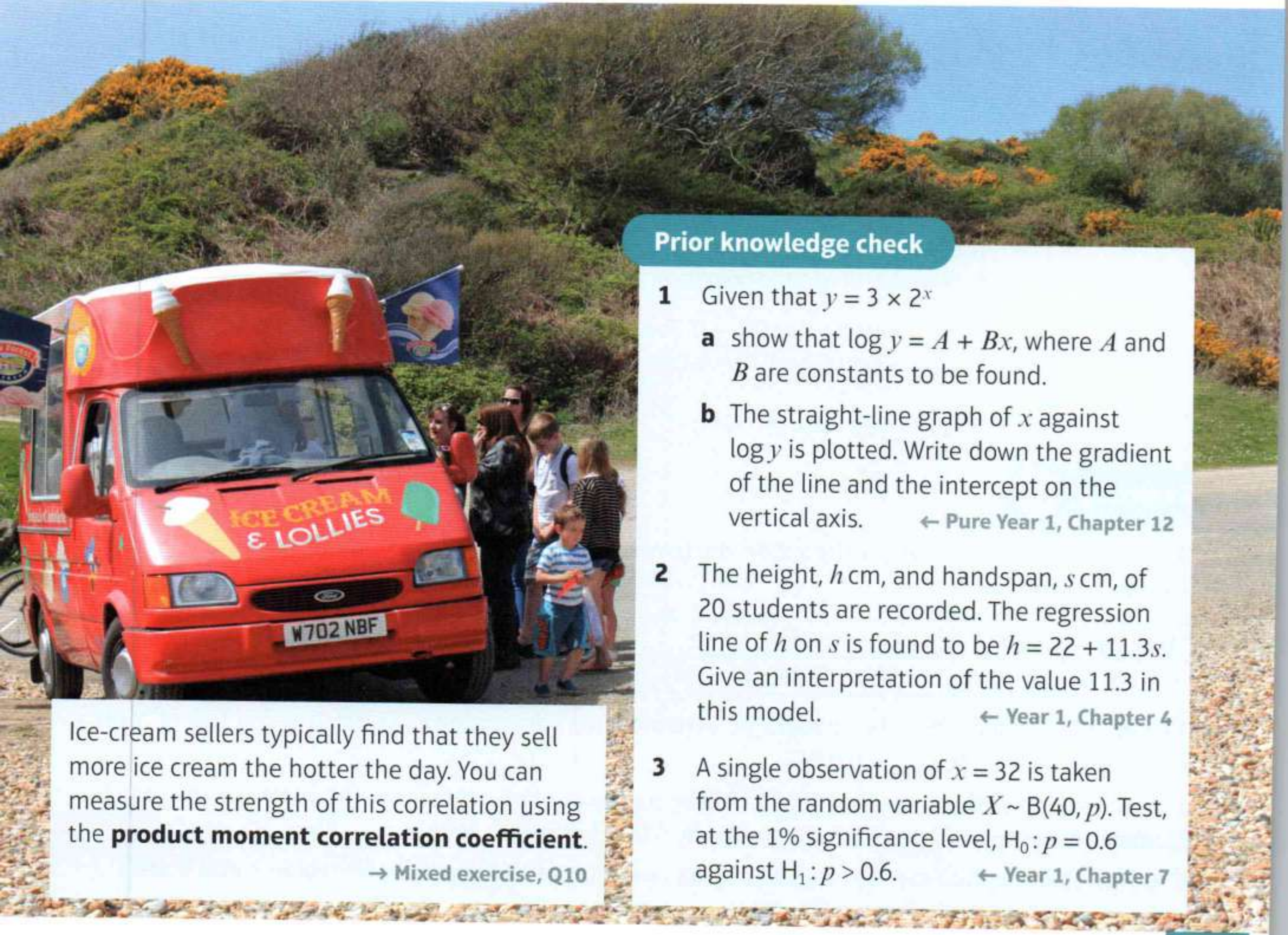
Regression, correlation and hypothesis testing 1

Objectives

After completing this chapter you should be able to:

- Understand exponential models in bivariate data → pages 2–5
- Use a change of variable to estimate coefficients in an exponential model → pages 2–5
- Understand and calculate the product moment correlation coefficient → pages 5–8
- Carry out a hypothesis test for zero correlation → pages 8–12



Ice-cream sellers typically find that they sell more ice cream the hotter the day. You can measure the strength of this correlation using the **product moment correlation coefficient**.

→ Mixed exercise, Q10

Prior knowledge check

1 Given that $y = 3 \times 2^x$

a show that $\log y = A + Bx$, where A and B are constants to be found.

b The straight-line graph of x against $\log y$ is plotted. Write down the gradient of the line and the intercept on the vertical axis. ← Pure Year 1, Chapter 12

2 The height, h cm, and handspan, s cm, of 20 students are recorded. The regression line of h on s is found to be $h = 22 + 11.3s$. Give an interpretation of the value 11.3 in this model. ← Year 1, Chapter 4

3 A single observation of $x = 32$ is taken from the random variable $X \sim B(40, p)$. Test, at the 1% significance level, $H_0 : p = 0.6$ against $H_1 : p > 0.6$. ← Year 1, Chapter 7

1.1 Exponential models

Regression lines can be used to model a **linear** relationship between two variables. Sometimes, experimental data does not fit a linear model, but still shows a clear pattern. You can use logarithms and coding to examine trends in non-linear data.

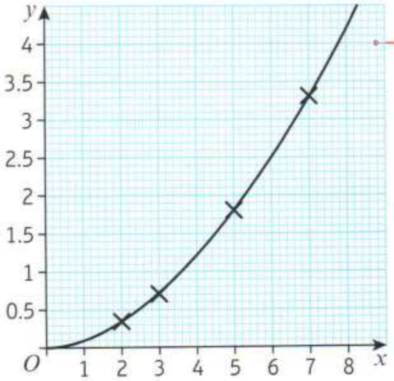
For data that can be modelled by a relationship of the form $y = ax^n$, you need to code the data using $Y = \log y$ and $X = \log x$ to obtain a linear relationship.

- **If $y = ax^n$ for constants a and n then $\log y = \log a + n \log x$**

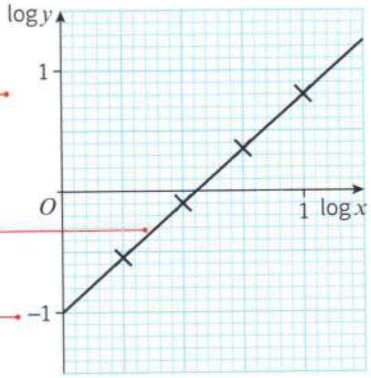
For data that can be modelled by an **exponential** relationship of the form $y = ab^x$, you need to code the data using $Y = \log y$ and $X = x$ to obtain a linear relationship.

- **If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$**

Link Take logs of both sides and rearrange to convert the original form into the linear form.
← Pure Year 1, Section 14.6



The points on this scatter graph satisfy the relationship $y = 0.1x^{1.8}$. This is in the form $y = ax^n$.



Plotting $\log x$ against $\log y$ gives a straight line.

The gradient of the line is 1.8. This corresponds to the value of n in the non-linear relationship.

The y-intercept is at (0, −1). This corresponds to $\log a$ hence $a = 10^{-1} = 0.1$, as expected.

Example 1

The table shows some data collected on the temperature, in °C, of a colony of bacteria (t) and its growth rate (g).

Temperature, t (°C)	3	5	6	8	9	11
Growth rate, g	1.04	1.49	1.79	2.58	3.1	4.46

The data are coded using the changes of variable $x = t$ and $y = \log g$. The regression line of y on x is found to be $y = -0.2215 + 0.0792x$.

a Mika says that the constant −0.2215 in the regression line means that the colony is shrinking when the temperature is 0 °C. Explain why Mika is wrong.

b Given that the data can be modelled by an equation of the form $g = kb^t$ where k and b are constants, find the values of k and b.

a When $t = 0$, $x = 0$, so according to the model,

$y = -0.2215$

$\log g = -0.2215$

$g = 10^{-0.2215} = 0.600$ (3 s.f.).

This growth rate is positive: the colony is not shrinking.

Remember that the original data have been coded. Use the coding in reverse to find the corresponding value of g. You could also observe that a prediction based on $t = 0$ would be outside the range of the data so would be an example of **extrapolation**. ← **Year 1, Section 4.2**

b Substitute $x = t$ and $y = \log g$:

$\log g = -0.2215 + 0.0792t$

$g = 10^{-0.2215 + 0.0792t}$

$g = 10^{-0.2215} \times (10^{0.0792})^t$

$g = 0.600 \times 1.20^t$ (both values given to 3 s.f.)

So $k = 0.600$ and $b = 1.20$

Use the change of variable to find an expression for $\log g$ in terms of t. You could also compare the equation of the regression line with $\log g = \log k + t \log b$. ← **Pure Year 1, Section 14.6**

Remember log means log to the base 10. So $10^{\log g} = g$.

Use the laws of indices to write the expression in the form $g = kb^t$.

Exercise 1A

Online Explore the original and coded data graphically using technology.

1 Data are coded using $Y = \log y$ and $X = \log x$ to give a linear relationship. The equation of the regression line for the coded data is $Y = 1.2 + 0.4X$.

a State whether the relationship between y and x is of the form $y = ax^n$ or $y = kb^x$.

b Write down the relationship between y and x and find the values of the constants.

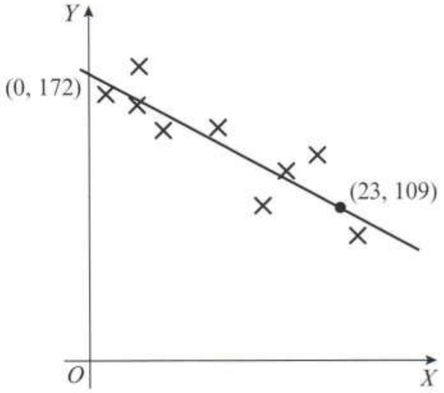
2 Data are coded using $Y = \log y$ and $X = x$ to give a linear relationship. The equation of the regression line for the coded data is $Y = 0.4 + 1.6X$.

a State whether the relationship between y and x is of the form $y = ax^n$ or $y = kb^x$.

b Write down the relationship between y and x and find the values of the constants.

(P) **3** The scatter diagram shows the relationship between two sets of coded data, X and Y, where $X = \log x$ and $Y = \log y$. The regression line of Y on X is shown, and passes through the points (0, 172) and (23, 109).

The relationship between the original data sets is modelled by an equation of the form $y = ax^n$. Find the exact value of a and the value of n correct to 3 decimal places.



(P) **4** The size of a population of moles is recorded and the data are shown in the table. T is the time, in months, elapsed since the beginning of the study and P is the number of moles in the population.

T	2	3	5	7	8	9
P	72	86	125	179	214	257

a Plot a scatter diagram showing log P against T.

b Comment on the correlation between log P and T.

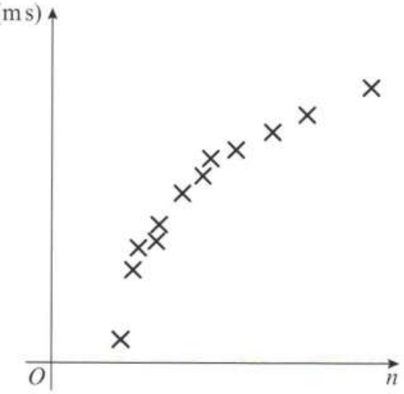
c State whether your answer to **b** supports the fact that the original data can be modelled by a relationship of the form $P = ab^T$.

d Approximate the values of a and b for this model.

e Give an interpretation of the value of b you calculated in part **d**.

Hint Think about what happens when the value of T increases by 1. When interpreting coefficients, refer in your answer to the context given in the question.

5 The time, t ms, needed for a computer algorithm to determine whether a number, n, is prime is recorded for different values of n. A scatter graph of t against n is drawn.



a Explain why a model of the form $t = a + bn$ is unlikely to fit these data.

The data are coded using the changes of variable $y = \log t$ and $x = \log n$. The regression line of y on x is found to be $y = -0.301 + 0.6x$.

b Find an equation for t in terms of n, giving your answer in the form $t = an^k$, where a and k are constants to be found.

6 Data are collected on the number of units (c) of a catalyst added to a chemical process, and the rate of reaction (r).

The data are coded using $x = \log c$ and $y = \log r$. It is found that a linear relationship exists between x and y and that the equation of the regression line of y on x is $y = 1.31x - 0.41$.

Use this equation to determine an expression for r in terms of c.

7 The heights, h cm, and masses, m kg, of a sample of Galapagos penguins are recorded. The data are coded using $y = \log m$ and $x = \log h$ and it is found that a linear relationship exists between x and y. The equation of the regression line of y on x is $y = 0.0023 + 1.8x$.

Find an equation to describe the relationship between m and h, giving your answer in the form $m = ah^n$, where a and n are constants to be found.

E/P **8** The table shows some data collected on the temperature, t °C, of a colony of insect larvae and the growth rate, g, of the population.

Temp, t (°C)	13	17	21	25	26	28
Growth rate, g	5.37	8.44	13.29	20.91	23.42	29.38

The data are coded using the changes of variable $x = t$ and $y = \log g$. The regression line of y on x is found to be $y = 0.09 + 0.05x$.

a Given that the data can be modelled by an equation of the form $g = ab^t$ where a and b are constants, find the values of a and b. **(3 marks)**

b Give an interpretation of the constant b in this equation. **(1 mark)**

c Explain why this model is not reliable for estimating the growth rate of the population when the temperature is 35 °C. **(1 mark)**

Challenge

The table shows some data collected on the efficiency rating, E, of a new type of super-cooled engine when operating at a certain temperature, T.

Temp, T (°C)	1.2	1.5	2	3	4	6	8
Efficiency, E	9	5.5	3	1.4	0.8	0.4	0.2

It is thought that the relationship between E and t is of the form $E = aT^b$.

a By plotting an appropriate scatter diagram, verify that this relationship is valid for the data given.

b By drawing a suitable line on your scatter diagram and finding its equation, estimate the values of a and b.

c Give a reason why the model will not predict the efficiency of the engine when the temperature is 0 °C.

1.2 Measuring correlation

You can calculate quantitative measures for the strength and type of linear correlation between two variables. One of these measures is known as the **product moment correlation coefficient**.

- **The product moment correlation coefficient describes the linear correlation between two variables. It can take values between −1 and 1.**

If $r = 1$, there is perfect positive linear correlation.

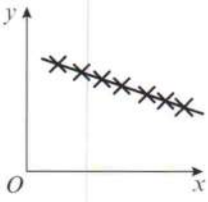
If $r = -1$, there is perfect negative linear correlation.

The closer r is to −1 or 1, the stronger the negative or positive correlation, respectively.

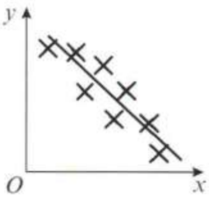
If $r = 0$ (or is close to 0) there is no linear correlation. In this case there might still be a non-linear relationship between the variables.

Notation The product moment correlation coefficient, or PMCC, for a sample of data is denoted by the letter r.

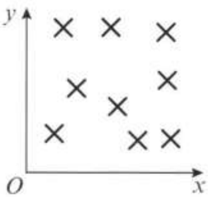
Hint For $r = \pm1$, the points all lie on a straight line.



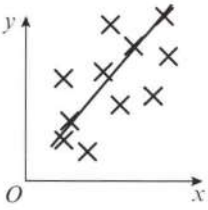
$r = -1$



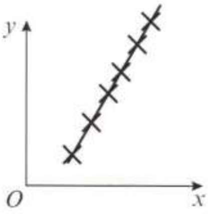
$r = -0.8$



$r = 0$



$r = 0.3$



$r = 1$

You need to know how to calculate the product moment correlation coefficient for bivariate data using your calculator.

Example 2

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 10 days in September in Hurn in 1987.

Day of month	1	2	3	4	5	6	7	8	9	10
w	4	4	8	7	12	12	3	4	7	10
g	13	12	19	23	33	37	10	n/a	n/a	23

© Crown Copyright Met Office

a State the meaning of n/a in the table above.

b Calculate the product moment correlation coefficient for the remaining 8 days.

c With reference to your answer to part **b**, comment on the suitability of a linear regression model for these data.

a Data on daily maximum gust is not available for these days.

b $r = 0.9533$

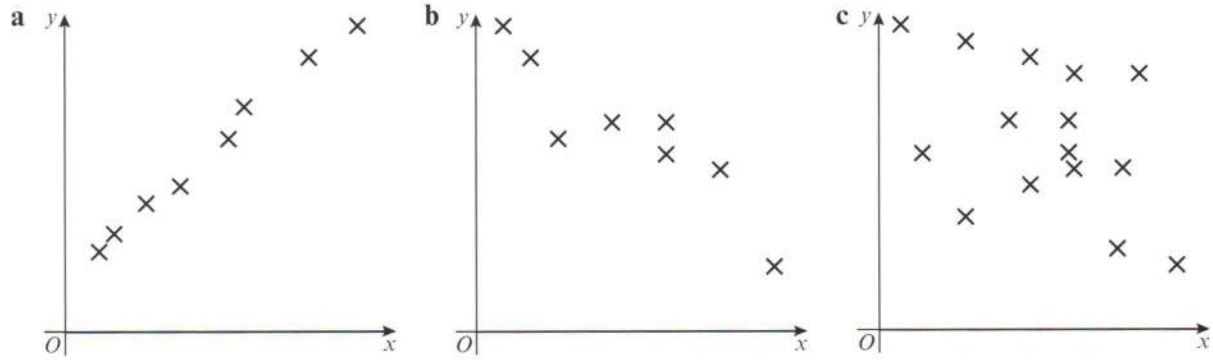
c r is close to 1 so there is a strong positive correlation between daily mean windspeed and daily maximum gust. This means that the data points lie close to a straight line, so a linear regression model is suitable.

Online Use your calculator to calculate the PMCC.

r measures linear correlation. The closer r is to 1 or −1, the more closely a linear regression model will fit the data.

Exercise 1B

1 Suggest a value of r for each of these scatter diagrams:



2 The following table shows 10 observations from a bivariate data set.

v	50	70	60	82	45	35	110	70	35	30
m	140	200	180	210	120	100	200	180	120	60

a State what is measured by the product moment correlation coefficient.

b Use your calculator to find the value of the product moment correlation coefficient between v and m.

3 In a training scheme for young people, the average time taken for each age group to reach a certain level of proficiency was measured. The table below shows the data.

Age, x (years)	16	17	18	19	20	21	22	23	24	25
Average time, y (hours)	12	11	10	9	11	8	9	7	6	8

a Use your calculator to find the value of the product moment correlation coefficient for these data.

b Use your answer to part **a** to describe the correlation between the age and average time taken based on this sample.

E/P **4** The number of atoms of a radioactive substance, n, is measured at various times, t minutes after the start of an experiment. The table below shows the data.

Time, t	1	2	4	5	7
Atoms, n	231	41	17	7	2
log n					

The data is coded using $x = t$ and $y = \log n$.

a Copy and complete the table showing the values of $\log n$. **(2 marks)**

b Calculate the product moment correlation coefficient for the coded data. **(1 mark)**

c With reference to your answer to **b**, state whether an exponential model is a good fit for these data. **(2 marks)**

The equation of the regression line of y on x is found to be $y = 2.487 - 0.320x$.

d Find an expression for n in terms of t, giving your answer in the form $n = ab^t$, where a and b are constants to be found. **(3 marks)**

Hint For part **b** enter corresponding values of t and $\log n$ into your calculator.

E/P **5** The width, w cm, and the mass, m grams, of snowballs are measured. The table below shows the data.

Width, w	3	4	6	8	11
Mass, m	23	40	80	147	265
log w					
log m					

The data are coded using $x = \log w$ and $y = \log m$.

a Copy and complete the table showing the values of $\log w$ and $\log m$. **(3 marks)**

b Calculate the product moment correlation coefficient for the coded data. **(1 mark)**

c With reference to your answer to **b**, state whether a model in the form $m = kw^n$ where k and n are constants is a good fit for these data. **(2 marks)**

The equation of the regression line of y on x is found to be $y = 0.464 + 1.88x$

d Determine the values of k and n. **(3 marks)**

E **6** From the large data set, the daily mean air temperature, t °C, and the rainfall, f mm, were recorded for Perth on seven consecutive days in August 2015.

Temp, t	18.0	16.4	15.3	15.0	13.7	10.2	12.0
Rainfall, f	3.0	13.0	4.6	32.0	28.0	63.0	22.0

© Crown Copyright Met Office

a Calculate the product moment correlation coefficient for these data. **(1 mark)**

b With reference to your answer to part **a**, comment on the suitability of a linear regression model for these data. **(2 marks)**

E/P **7** From the large data set, the daily total rainfall, x mm, and the daily total sunshine, y hours, were recorded for Camborne on seven consecutive days in May 2015.

Rainfall, x	2.2	tr	1.4	4.4	tr	0.2	0.6
Sunshine, y	5.2	7.7	5.6	0.3	5.1	0.1	8.9

© Crown Copyright Met Office

a State the meaning of 'tr' in the table above. **(1 mark)**

b Calculate the product moment correlation coefficient for these 7 days, stating clearly how you deal with the entries marked 'tr'. **(2 marks)**

c With reference to your answer to part **b**, comment on the suitability of a linear regression model for these data. **(2 marks)**

Challenge

Data are recorded for two variables, x and y.

x	3.1	5.6	7.1	8.6	9.4	10.7
y	3.2	4.8	5.7	6.5	6.9	7.6

By calculating the product moment correlation coefficients for suitably coded values of x and y state, with reasons, whether these data are more closely modelled by a relationship of the form $y = ab^x$ or a relationship of the form $y = kx^n$, where a, b, k and n are constants.

1.3 Hypothesis testing for zero correlation

You can use a hypothesis test to determine whether the product moment correlation coefficient, r, for a particular sample indicates that there is likely to be a linear relationship within the whole population.

Notation r denotes the PMCC for a **sample**. ρ denotes the PMCC for a **whole population**. It is the Greek letter *rho*.

If you want to test for whether or not the population PMCC, ρ, is either greater than zero or less than zero you can use a **one-tailed test**:

For a one-tailed test use either:

- $\mathbf{H_0 : \rho = 0, H_1 : \rho > 0}$ **or**
- $\mathbf{H_0 : \rho = 0, H_1 : \rho < 0}$

If you want to test whether the population PMCC, ρ, is not equal to zero you need to use a **two-tailed test**:

For a two-tailed test use:

- $\mathbf{H_0 : \rho = 0, H_1 : \rho \neq 0}$

You can determine the critical region for r for your hypothesis test by using the table of critical values on page 191. This table will be given in the *Mathematical Formulae and Statistical Tables* booklet in your exam. The critical region depends on the **significance level** of the test and the **sample size**.

Product moment coefficient					
		Level			**Sample**
0.10	**0.05**	**0.025**	**0.01**	**0.005**	**size**
0.8000	0.9000	0.9500	0.9800	0.9900	4
0.6870	0.8054	0.8783	0.9343	0.9587	5
0.6084	0.7293	0.8114	0.8822	0.9172	6
0.5509	0.6694	0.7545	0.8329	0.8745	7
0.5067	0.6215	0.7067	0.7887	0.8343	8
0.4716	0.5822	0.6664	0.7498	0.7977	9

For a sample size of 8 you see from the table that the critical value of r to be significant at the 5% level on a one-tailed test is 0.6215. An observed value of r greater than 0.6215 from a sample of size 8 would provide sufficient evidence to reject the null hypothesis and conclude that $\rho > 0$. Similarly, an observed value of r less than −0.6215 would provide sufficient evidence to conclude that $\rho < 0$.

Example 3

A scientist takes 30 observations of the masses of two reactants in an experiment. She calculates a product moment correlation coefficient of $r = -0.45$.

The scientist believes there is no correlation between the masses of the two reactants. Test, at the 10% level of significance, the scientist's claim, stating your hypotheses clearly.

$H_0 : \rho = 0, H_1 : \rho \neq 0$

You need to test for either positive or negative correlation, so use a **two-tailed** test.

Sample size = 30

Significance level in each tail = 0.05

Halve the significance level to find the probability in each tail. ← **Year 1, Section 7.4**

From the table, critical values of r for a 5% significance level with a sample size of 30 are $r = \pm 0.3061$, so the critical region is $r < -0.3061$ and $r > 0.3061$.

Use the table of critical values on page 191 to find the critical region for a two-tailed test with a total significance level of 10%.

$-0.45 < -0.3061$. The observed value of r lies within the critical region, so reject H_0.

You reject H_0 if the observed value lies inside the critical region. ← **Year 1, Section 7.2**

There is evidence, at the 10% level of significance, that there is a correlation between the masses of the two reactants.

Write a conclusion in the context of the original question.

Example 4

The table from the large data set shows the daily maximum gust, x kn, and the daily maximum relative humidity, y%, in Leeming for a sample of eight days in May 2015.

x	31	28	38	37	18	17	21	29
y	99	94	87	80	80	89	84	86

© Crown Copyright Met Office

a Find the product moment correlation coefficient for these data.

b Test, at the 10% level of significance, whether there is evidence of a positive correlation between daily maximum gust and daily maximum relative humidity. State your hypotheses clearly.

a $r = 0.1149$

Use your calculator.

b $H_0 : \rho = 0$, $H_1 : \rho > 0$

Sample size = 8

Significance level = 0.1

You are testing for evidence of **positive** correlation, so use a **one-tailed test**.

From the table, the critical value of r is 0.5067 and the critical region is $r > 0.5067$

Use the table of critical values on page 191 to find the critical region for a one-tailed test with a significance level of 10%.

$0.1149 < 0.5067$. The observed value of r is not in the critical region, so there is not enough evidence to reject H_0.

If the observed value does not lie inside the critical region, you do not reject the null hypothesis. ← **Year 1, Section 7.2**

There is not sufficient evidence, at the 10% level of significance, of a positive correlation between the daily maximum gust and the daily maximum relative humidity.

Exercise 1C

1 A population of students each took two different tests. A sample of 40 students was taken from the population and their scores on the two tests were recorded. A product moment correlation coefficient of 0.3275 was calculated. Test whether or not this shows evidence of correlation between the test scores:

a at the 5% level

b at the 2% level.

Hint 'Evidence of correlation' could mean either positive or negative correlation, so you need to use a two-tailed test with $H_0 : \rho = 0$, $H_1 : \rho \neq 0$

2 A computer-controlled milling machine is calibrated between 1 and 7 times a week. A supervisor recorded the number of weekly calibrations, x, and the number of manufacturing errors, y, in each of 7 weeks.

x	2	3	1	7	6	5	4
y	53	55	62	19	35	40	41

a Calculate the product moment correlation coefficient for these data.

b For these data, test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$, using a 1% significance level.

E/P **3** **a** State what is measured by the product moment correlation coefficient. **(1 mark)**

Twelve students sat two Biology tests, one theoretical the other practical. Their marks are shown below.

Marks in theoretical test, *t*	5	9	7	11	20	4	6	17	12	10	15	16
Marks in practical test, *p*	6	8	9	13	20	9	8	17	14	8	17	18

b Find the product moment correlation coefficient for these data, correct to 3 significant figures. **(2 marks)**

A teacher claims that students who do well in their theoretical test also tend to do well in their practical test.

c Test this claim at a 0.05 significance level, stating your hypotheses clearly. **(3 marks)**

d Give an interpretation of the value 0.05 in your hypothesis test. **(1 mark)**

E **4** The following table shows the marks attained by 8 students in English and Mathematics tests.

Student	A	B	C	D	E	F	G	H
English	25	18	32	27	21	35	28	30
Mathematics	16	11	20	17	15	26	32	20

a Calculate the product moment correlation coefficient. **(1 mark)**

b Test, at the 5% significance level, whether these results show evidence of a positive linear relationship between English and Mathematics marks. State your hypotheses clearly. **(3 marks)**

5 A small company decided to import fine Chinese porcelain. They believed that in the long term this would prove to be an increasingly profitable arrangement with profits increasing proportionally to sales. Over the next 6 years their sales and profits were as shown in the table below.

Year	1994	1995	1996	1997	1998	1999
Sale in thousands	165	165	170	178	178	175
Profits in £1000	65	72	75	76	80	83

Using a 1% significance level, test to see if there is any evidence that the company's beliefs were correct, and that profit and sales were positively correlated.

E **6** A scientific researcher collects data on the amount of solvent in a solution and the rate of reaction. She calculates the product moment correlation coefficient between the two sets of data and finds it to be −0.43. Given that she collected data from 15 samples, test, at the 5% level of significance, the claim that there is a negative correlation between the amount of solvent and the rate of reaction. State your hypotheses clearly. **(3 marks)**

P **7** A safari ranger believes that there is a positive correlation between the amount of grass per square kilometre and the number of meerkats that graze there. He decides to carry out a hypothesis test to see if there is evidence for his claim. He takes a random sample of 10 equal-sized areas of grassland, records the amount of grass and the number of meerkats grazing in each, and finds that the correlation coefficient is 0.66.

Given that this result provided the ranger with sufficient evidence to reject his null hypothesis, suggest the least possible significance level for the ranger's test.

Ⓟ **8** Data on the daily mean temperature and the daily total sunshine is taken from the large data set for Leuchars in May and June 1987. A meteorologist finds that the product moment correlation coefficient for these data is 0.715. Given that the researcher tests for positive correlation at the 2.5% level of significance, and concludes that the value is significant, find the smallest possible sample size.

Ⓔ **9** An employee at a weather centre believes that there is a negative correlation between humidity and visibility. She takes a sample of data from Heathrow in August 1987.

Humidity (%)	92	93	91	82	91	100
Visibility (m)	2500	1500	2700	2900	2200	1000

© Crown Copyright Met Office

a Calculate the product moment correlation coefficient for these data. **(1 mark)**

b Test, at the 1% level of significance, the employee's claim. State your hypotheses clearly. **(3 marks)**

Ⓟ **10** A scientist wishes to test, at the 5% level, whether there is any correlation between the masses of two reactants in an experiment. She conducts the experiment 20 times and observes a product moment correlation coefficient of $r = 0.4$.

The probability of obtaining this value of r or higher, given $\rho = 0$, is 0.0403. The scientist claims that this means her result is significant at the 5% level.

a Explain why the scientist is incorrect.

b Find the critical values of r for this test at the 5% level.

Mixed exercise 1

Ⓔ **1** Conor uses a 3D printer to produce various pieces for a model. He records the time taken, t hours, to produce each piece, and its base area, $x\,\text{cm}^2$.

Base area, x (cm^2)	1.1	1.3	1.9	2.2	2.5	3.7
Time, t (hours)	0.7	0.9	1.5	1.8	2.2	3.8

a Calculate the product moment correlation coefficient between $\log x$ and $\log t$. **(2 marks)**

b Use your answer to part **a** to explain why an equation of the form $t = ax^n$, where a and n are constants, is likely to be a good model for the relationship between x and t. **(1 mark)**

c The regression line of $\log t$ on $\log x$ is given as $\log t = -0.210 + 1.38 \log x$. Determine the values of the constants a and n in the equation given in part **b**. **(2 marks)**

Ⓔ **2** The table shows some data collected on the temperature in °C of a chemical reaction (t) and the amount of dry residue produced (d grams).

Temperature, t (°C)	38	51	72	83	89	94
Dry residue, d (grams)	4.3	11.7	58.6	136.7	217.0	318.8

The data are coded using the changes of variable $x = t$ and $y = \log d$. The regression line of y on x is found to be $y = -0.635 + 0.0334x$.

a Given that the data can be modelled by an equation of the form $d = ab^t$ where a and b are constants, find the values of a and b. **(3 marks)**

b Explain why this model is not reliable for estimating the amount of dry residue produced when the temperature is 151 °C. **(1 mark)**

3 The product moment correlation coefficient for a person's age and his score on a memory test is –0.86. Interpret this value.

P **4** Each of 10 cows was given an additive (x) every day for four weeks to see if it would improve the milk yield (y). At the beginning, the average milk yield per day was 4 gallons. The milk yield of each cow was measured on the last day of the four weeks. The table shows the data.

Cow	A	B	C	D	E	F	G	H	I	J
Additive, x (25 g)	1	2	3	4	5	6	7	8	9	10
Yield, y (gallons)	4.0	4.2	4.3	4.5	4.5	4.7	5.2	5.2	5.1	5.1

a By drawing a scatter diagram or otherwise, suggest the maximum amount of additive that should be given to the cows to maximise yield.

b Calculate the value of the product moment correlation coefficient for the first seven cows.

c Without further calculation, write down, with a reason, how the product moment correlation coefficient for all 10 cows would differ from your answer to **b**.

E **5** The following table shows the engine size (c), in cubic centimetres, and the fuel consumption (f), in miles per gallon to the nearest mile, for 10 car models.

c (cm^3)	1000	1200	1400	1500	1600	1800	2000	2200	2500	3000
f (mpg)	46	42	43	39	41	37	35	29	28	25

a Use your calculator to find the value of the product moment correlation coefficient between c and f. **(1 mark)**

b Interpret your answer to part **a**. **(2 marks)**

E **6** As part of a survey in a particular profession, age, x years, and yearly salary, £y thousands, were recorded. The values of x and y for a randomly selected sample of ten members of the profession are as follows:

x	30	52	38	48	56	44	41	25	32	27
y	22	38	40	34	35	32	28	27	29	41

a Calculate, to 3 decimal places, the product moment correlation coefficient between age and salary. **(1 mark)**

It is suggested that there is no correlation between age and salary.

b Test this suggestion at the 5% significance level, stating your null and alternative hypotheses clearly. **(3 marks)**

E **7** A machine hire company kept records of the age, X months, and the maintenance costs, £Y, of one type of machine. The table summarises the data for a random sample of 10 machines.

Machine	A	B	C	D	E	F	G	H	I	J
Age, X	63	12	34	81	51	14	45	74	24	89
Maintenance costs, Y	111	25	41	181	64	21	51	145	43	241

a Calculate, to 3 decimal places, the product moment correlation coefficient. **(1 mark)**

It is believed that there is a relationship between the age and maintenance cost of these machines.

b Using a 5% level of significance and quoting from the table of critical values, interpret your correlation coefficient. Use a two-tailed test and state clearly your null and alternative hypotheses. **(3 marks)**

E **8** The data below show the height above sea level, x metres, and the temperature, y °C, at 7.00 a.m., on the same day in summer at nine places in Europe.

Height, x (m)	1400	400	280	790	390	590	540	1250	680
Temperature, y (°C)	6	15	18	10	16	14	13	7	13

The product moment correlation coefficient is −0.975. Use this value to test for negative correlation at the 5% significance level. Interpret your result in context. **(3 marks)**

E **9** The ages, in months, and the weights, in kg, of a random sample of nine babies are shown in the table below.

Baby	A	B	C	D	E	F	G	H	I
Age, x	1	2	2	3	3	3	4	4	5
Weight, y	4.4	5.2	5.8	6.4	6.7	7.2	7.6	7.9	8.4

The product moment correlation coefficient between weight and age for these babies was found to be 0.972. By testing for positive correlation at the 5% significance level interpret this value. **(3 marks)**

E **10** An ice-cream seller believes that there is a positive correlation between the amount of sunshine and sales of ice cream. He collects data on six days during June 2015 at his 'pitch' in Camborne:

Sunshine (hours)	4.2	7.9	13.8	8.7	6.2	0.7
Ice-cream sales (£100s)	7.0	8.3	12.4	8.1	7.9	6.2

a Calculate the product moment correlation coefficient for these data. **(1 mark)**

b Carry out a hypothesis test to determine, at the 5% level, if there is significant evidence in support of the ice-cream seller's belief. State your hypotheses clearly. **(3 marks)**

E **11** A meteorologist believes that there is a positive correlation between daily mean windspeed and daily maximum gust. She collects data from the large data set for 5 days during August 2015 in the town of Hurn.

Mean windspeed (knots)	4	7	7	8	5
Daily maximum gust (knots)	14	22	18	20	17

© Crown Copyright Met Office

By calculating the product moment correlation coefficient for these data, test at the 5% level of significance whether there is evidence to support the meteorologist's claim. State your hypotheses clearly. **(4 marks)**

Ⓔ **12** The table shows data from the large data set on the daily mean air temperature and the daily mean pressure during May and June 2015 in Beijing.

Temperature (°C)	17.5	18.5	18.0	24.6	22.2	23.1	27.3
Pressure (hPa)	1010	1011	1012	997	1009	998	1002

© Crown Copyright Met Office

Test at the 2.5% level of significance the claim that there is negative correlation between the daily mean air temperature and the daily mean pressure. State your hypotheses clearly.

(4 marks)

Large data set

You will need access to the large data set and spreadsheet software to answer these questions.

1 a Take a random sample of size 20 from the data for Heathrow in 2015, and record the daily mean air temperature and daily total rainfall.

b Calculate the product moment correlation coefficient between these variables for your sample.

c Test, at the 5% level of significance, the claim that there is a correlation between the daily mean air temperature and the daily total rainfall.

2 a State with a reason whether you would expect to find a relationship between daily mean total cloud cover and daily mean visibility.

b Use a random sample from the large data set to test for this relationship. You should state clearly:

- Your sample size and location
- Your sampling method
- The hypotheses and significance level for your test
- A conclusion in the context of the question

Hint You might be able to use the **Correl** or **CorrelationCoefficient** commands in your spreadsheet software to calculate the PMCC.

Summary of key points

1 If $y = ax^n$ for constants a and n then $\log y = \log a + n \log x$

2 If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$

3 The **product moment correlation coefficient** describes the linear correlation between two variables. It can take values between −1 and 1.

4 For a one-tailed test use either:

- $\mathrm{H}_0: \rho = 0, \mathrm{H}_1: \rho > 0$ or
- $\mathrm{H}_0: \rho = 0, \mathrm{H}_1: \rho < 0$

For a two-tailed test use:

- $\mathrm{H}_0: \rho = 0, \mathrm{H}_1: \rho \neq 0$