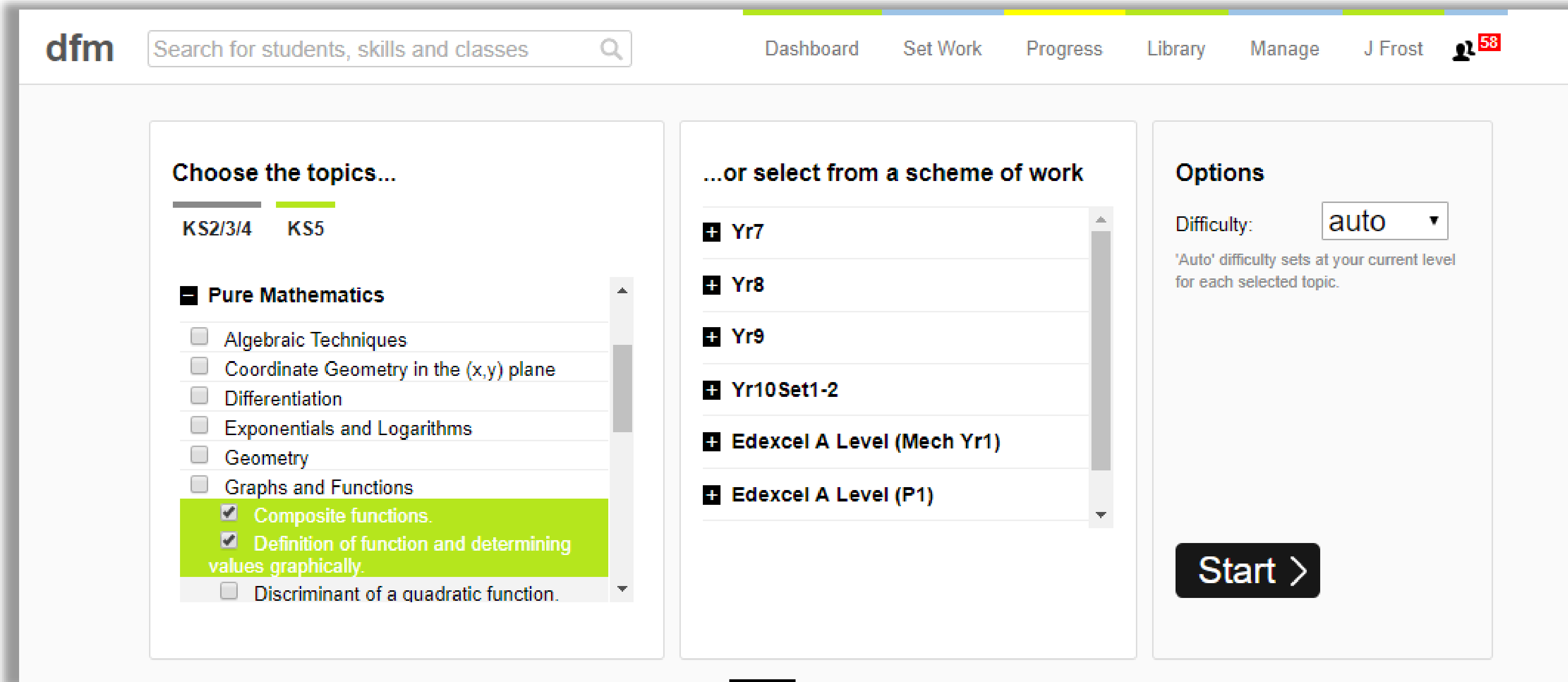
Stats Yr2 Chapter 1 :: Regression, Correlation & Hypothesis Tests

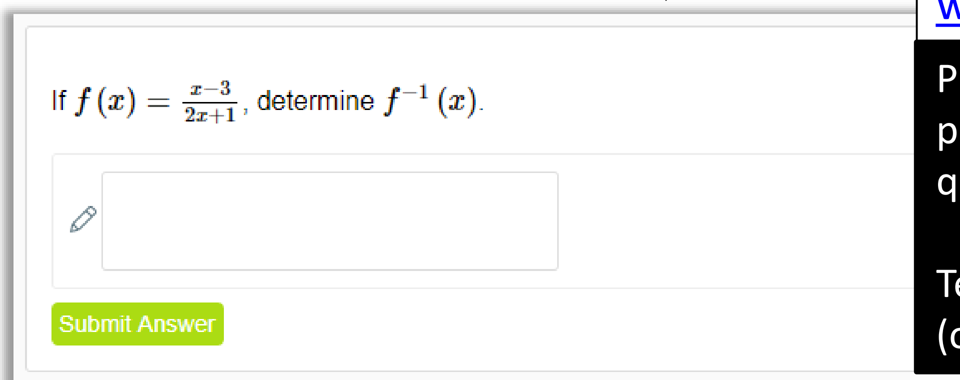
jfrost@tiffin.kingston.sch.uk

www.drfrostmaths.com
@DrFrostMaths

Last modified: 22nd January 2018

Use of DrFrostMaths for practice





Register for **free** at:
www.drfrostmaths.com/homework

Practise questions by chapter, including past paper Edexcel questions and extension questions (e.g. MAT).

Teachers: you can create student accounts (or students can register themselves).

Chapter Overview

1:: Exponential Models

Recap of Pure Year 1. Using $y = ab^x$ to model an exponential relationship between two variables.

2:: Measuring Correlation

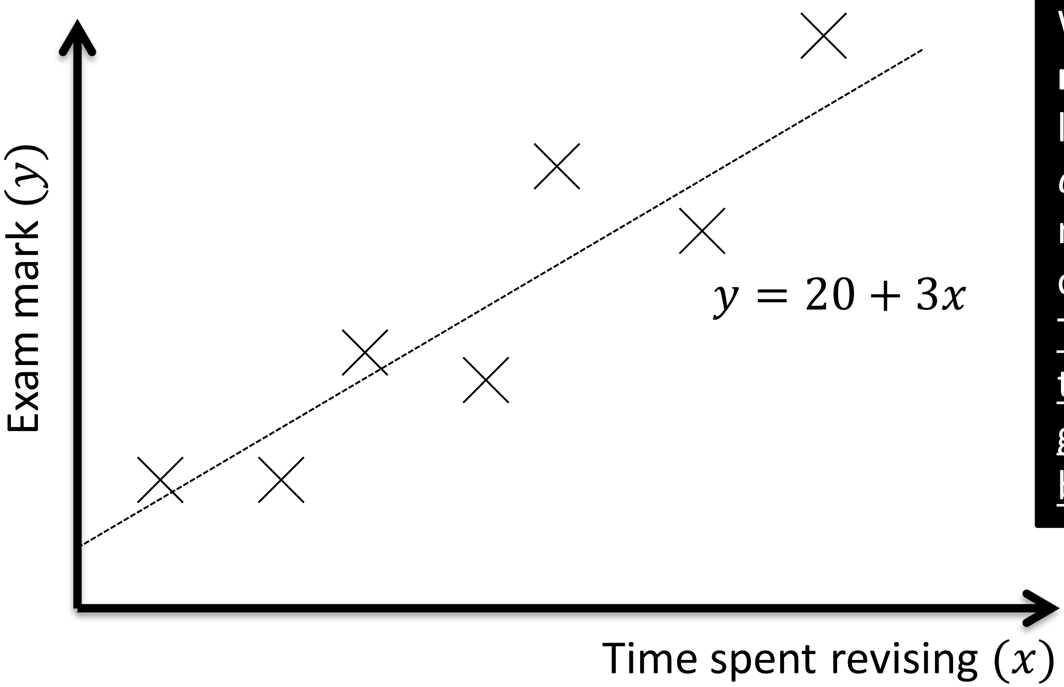
Using the Product Moment Correlation Coefficient (PMCC), r, to measure the strength of correlation between two variables.

3:: Hypothesis Testing for no correlation

We want to test whether two variables have some kind of correlation, or whether any correlation observed just happened by chance.

Teacher Notes: (1) is mostly a recap of Pure Year 1. (2) is in the old S1 module, but students now just use their calculator to calculate r; they do not need to use formulae. (3) is from the old S3 module but simplified.

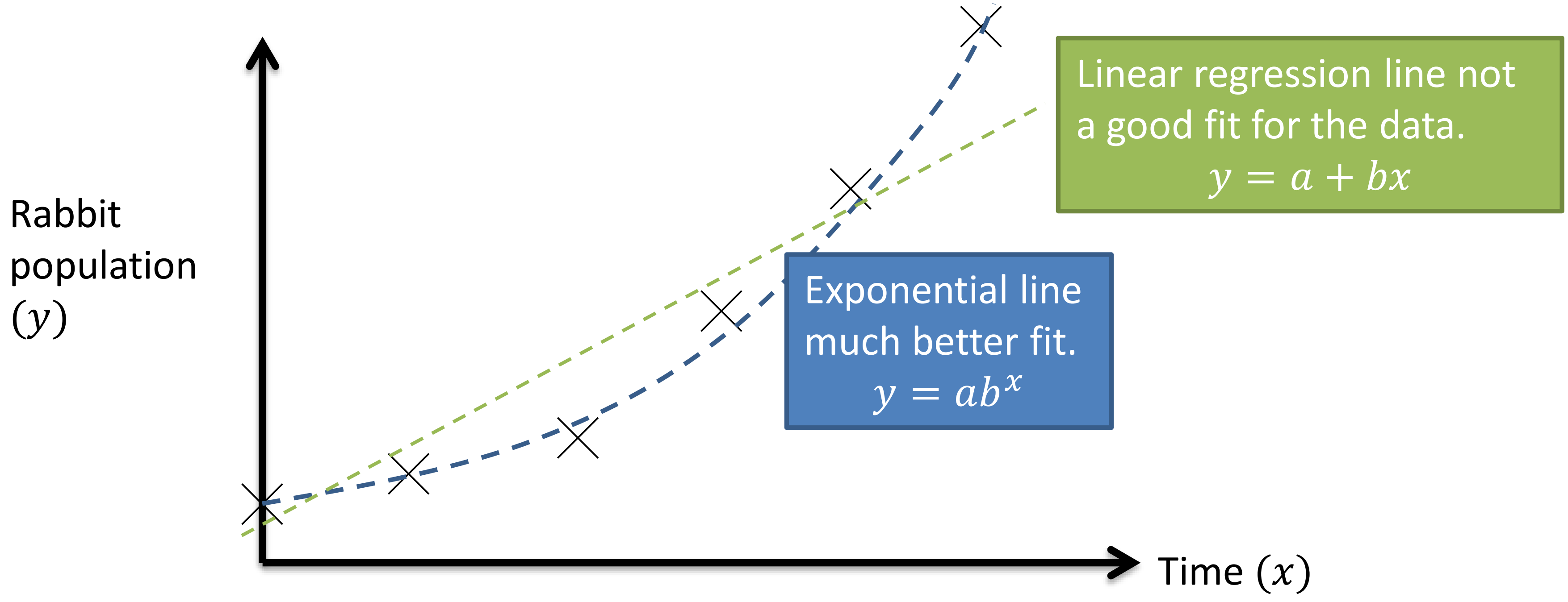
RECAP :: What is regression?



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.
<u>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.</u>

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. How would I do this?

Exponential Regression



For some variables, e.g. population with time, it may be more appropriate to use an **exponential** equation, i.e. $y = ab^x$, where a and b are constants we need to fix to best match the data.

$y = ab^x$

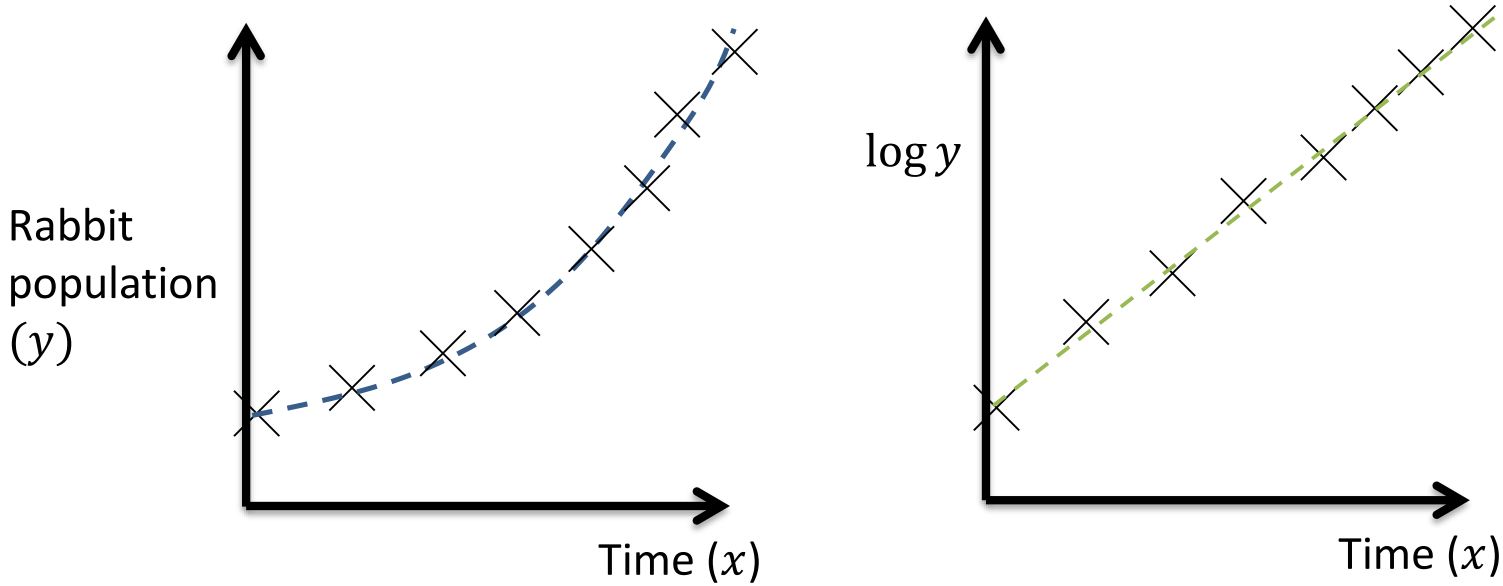
?

In Year 1, what did we do to both sides to end up with a straight line equation?

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

Exponential Regression

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



Comparing the equations, we can see that if we log the y values (although leave the x values), the data then forms a straight line, with y-intercept $\log k$ and gradient $\log b$.

Example

[Textbook] 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

?

b

?

Test Your Understanding

Robert wants to model a rabbit population P with respect to time in years t. He proposes that the population can be modelled using an exponential model: $P = kb^t$ The data is coded using $x = t$ and $y = \log P$. The regression line of y on x is found to be $y = 2 + 0.3x$. Determine the values of k and b.

?

Rabbit

Exercise 1A

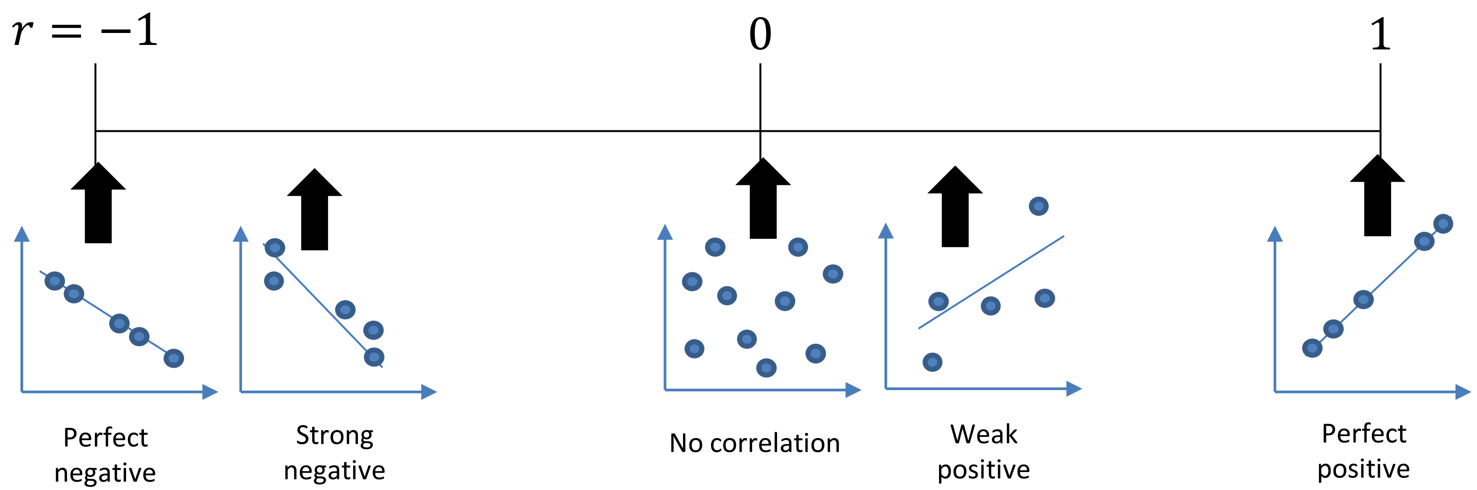
Pearson Pure Mathematics Year 1/AS

Pages 3-5

Measuring Correlation

You're used to use qualitative terms such as "positive correlation" and "negative correlation" and "no correlation" to describe the **type** of correlation, and terms such as "perfect", "strong" and "weak" to describe the **strength**.
The **Product Moment Correlation Coefficient** is one way to quantify this:

✎ The product moment correlation coefficient (PMCC), denoted by r, describes the linear correlation between two variables. It can take values between -1 and 1.



Rule of thumb: $r < -0.7$ or $r > 0.7$ is considered to be 'strong' correlation.

Note that PMCC is only applicable for a linear correlation, i.e. closeness of fit to a linear regression line (i.e. a straight 'line of best fit'). It may be the data exhibits strong correlation with respect to a different model (e.g. exponential) even when the PMCC is low.

Calculating r on your calculator

You must have a calculator that is capable of calculating r directly: in the A Level 2017+ syllabus you are no longer required to use formulae to calculate r.

x	y
1	3
2	6
3	5
4	8

6: Statistics

The following instructions are for the Casio ClassWiz. Press MODE then select 'Statistics'.

$y = a + bx$

We want to measure **linear** correlation, so select $y = a + bx$

Data Entry

Enter each of the x values in the table on the left, press = after each input. Use the arrow keys to get to the top of the y column.

PMCC

While entering data, press OPTN then choose "Regression Calc" to obtain r (i.e. the coefficients of your line of best fit and the PMCC). a and b would give you the y-intercept and gradient of the regression line (but not required in this chapter).

Pressing AC allows you to construct a statistical calculation yourself. In OPTN, there is an additional 'Regression' menu allowing you to insert r into your calculation.

You should obtain $r = 0.868$

Example

[Textbook] 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

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 ?

b ?

c ?

This is a common exam question. The important bit is **evaluating the suitability of the chosen model** (in this case a linear regression model, i.e. line of best fit). The closer r is to 1 or to -1, the more suitable this linear regression model.

Exercise 1B

Pearson Pure Mathematics Year 1/AS
Pages 6-8

Hypothesis Testing for correlation

Student	English Exam Mark	Maths Exam Mark
	Mean 60	Mean 70
	S.D. 5	S.D. 10
1	63.90	70.13
2	55.24	65.99
3	58.80	80.18
4	59.65	57.16
5	66.44	72.76
6	59.53	79.82
7	57.43	71.48
8	58.33	60.56
9	67.43	69.56
10	63.11	87.13
	r= 0.219	

Student	English Exam Mark	Maths Exam Mark
	Mean 60	Mean 70
	S.D. 5	S.D. 10
1	60.22	74.64
2	62.25	79.15
3	61.30	75.29
4	60.61	71.35
5	55.31	74.05
6	57.13	89.73
7	57.16	70.41
8	58.96	60.31
9	56.30	71.95
10	63.23	69.95
	r= -0.094	

Suppose we use a spreadsheet to randomly generate maths marks for students, and separately generate random English marks.
(This Excel demo accompanies this file – you can press F9 in Excel to generate a new set of random data)

What is the **observed** PMCC between Maths and English marks in this first set of data?

?

But what is the true underlying PMCC between Maths and English?

?

✎ r denotes the PMCC of a **sample**.

✎ ρ (Greek letter *rho*) is the PMCC for the **whole population**.

✎ $\therefore$ r is the test statistic, ρ is the population parameter.

Hypothesis Testing for correlation

	English Exam Mark		Maths Exam Mark	
	Mean	60	Mean	70
Student	S.D.	5	S.D.	10
1		63.90		70.13
2		55.24		65.99
3		58.80		80.18
4		59.65		57.16
5		66.44		72.76
6		59.53		79.82
7		57.43		71.48
8		58.33		60.56
9		67.43		69.56
10		63.11		87.13
	r=	0.219		

	English Exam Mark		Maths Exam Mark	
	Mean	60	Mean	70
Student	S.D.	5	S.D.	10
1		60.22		74.64
2		62.25		79.15
3		61.30		75.29
4		60.61		71.35
5		55.31		74.05
6		57.13		89.73
7		57.16		70.41
8		58.96		60.31
9		56.30		71.95
10		63.23		69.95
	r=	-0.094		

So despite that $\rho = 0$ (i.e. English/Maths marks have no underlying correlation), the observed PMCC for the sample, r, could vary, and might, purely by chance, suggest some correlation. In the first sample, this happened to be positive, in the latter, negative.

If we increased the sample size (in this case 10), we'd expect the observed PMCC to more closely match the true PMCC (0).

A natural question we might therefore ask is, if we had no knowledge about the underlying population: "**is the PMCC of $r = 0.219$ statistically significant, or are Maths/English marks not correlated, and this observed correlation emerged just by chance**?"

This sounds like a Hypothesis Test!

How to carry out the hypothesis test

	B	C	D	E	G	H
1		English Exam Mark			Maths Exam Mark	
2		Mean	60		Mean	70
3	Student	S.D.	5		S.D.	10
4	1		63.90			70.13
5	2		55.24			65.99
6	3		58.80			80.18
7	4		59.65			57.16
	5		66.44			72.76
	6		59.53			79.82
10	7		57.43			71.48
11	8		58.33			60.56
12	9		67.43			69.56
13	10		63.11			87.13
16		r=	0.219			

Let's carry out a hypothesis test on whether there is positive correlation between English and Maths marks, at 5% significance level:

H_0: ?

H_1: ?

There is no underlying correlation between English and maths marks.

There is an underlying positive correlation between English and maths marks.

Sample size

?

Critical value for 5% significance level:

?

Look up value in table.

?

As with Year 1, a 2 mark conclusion: (a) Compare values; do we reject H_0? (b) put in context of original problem.

CRITICAL VALUES FOR CORRELATION COEFFICIENTS

These tables concern tests of the hypothesis that a population correlation coefficient ρ is 0. The values in the tables are the minimum values which need to be reached by a sample correlation coefficient in order to be significant at the level shown, on a one-tailed test.

Product Moment Coefficient					Sample	Spearman's Coefficient		
Level						Level		
0.10	0.05	0.025	0.01	0.005	Level	0.05	0.025	0.01
0.8000	0.9000	0.9500	0.9800	0.9900	4	1.0000	-	-
0.6870	0.8054	0.8783	0.9343	0.9587	5	0.9000	1.0000	1.0000
0.6084	0.7293	0.8114	0.8822	0.9172	6	0.8286	0.8857	0.9429
0.5509	0.6694	0.7545	0.8329	0.8745	7	0.7143	0.7857	0.8929
0.5067	0.6215	0.7067	0.7887	0.8343	8	0.6429	0.7381	0.8333
0.4716	0.5822	0.6664	0.7498	0.7977	9	0.6000	0.7000	0.7833
0.4428	0.5494	0.6319	0.7155	0.7646	10	0.5636	0.6485	0.7455
0.4187	0.5214	0.6021	0.6851	0.7348	11	0.5364	0.6182	0.7091
0.3981	0.4973	0.5760	0.6581	0.7079	12	0.5035	0.5874	0.6783
0.3802	0.4762	0.5529	0.6339	0.6835	13	0.4835	0.5604	0.6484
0.3646	0.4575	0.5324	0.6120	0.6614	14	0.4637	0.5385	0.6264

These values give the minimum value of r required to reject the null hypothesis, i.e. the amount of correlation that would be considered significant.

Two-tailed test

In the previous example we hypothesised that English/Maths marks were positively correlated. But we could also test whether there was **any** correlation, i.e. positive **or** negative.

[Textbook] 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.

Product Moment Coefficient					
Level					Sample size, n
0.10	0.05	0.025	0.01	0.005	
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
0.2992	0.3783	0.4438	0.5155	0.5614	20
0.2914	0.3687	0.4329	0.5034	0.5487	21
0.2841	0.3598	0.4227	0.4921	0.5368	22
0.2774	0.3515	0.4133	0.4815	0.5256	23
0.2711	0.3438	0.4044	0.4716	0.5151	24
0.2653	0.3365	0.3961	0.4622	0.5052	25
0.2598	0.3297	0.3882	0.4534	0.4958	26
0.2546	0.3233	0.3809	0.4451	0.4869	27
0.2497	0.3172	0.3739	0.4372	0.4785	28
0.2451	0.3115	0.3673	0.4297	0.4705	29
0.2407	0.3061	0.3610	0.4226	0.4629	30
0.2070	0.2638	0.3120	0.3665	0.4026	40
0.1843	0.2353	0.2787	0.3281	0.3610	50
0.1678	0.2144	0.2542	0.2997	0.3301	60

H_0: ?

H_1: ?

Sample size = ?

Critical value at ? significance:

?

Two-tailed, so 5% at each tail.

?

Test Your Understanding

[Textbook] 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

a. Find the product moment correlation coefficient for this 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 ?

b ?

Exercise 1C

Pearson Pure Mathematics Year 1/AS
Pages 10-12